Reconstructing the Range Expansion of the European Starling in Southern Africa using a Hybrid Method of Niche Modelling and Individual Based Modelling

by

Marinel Janse van Rensburg

Thesis presented in partial fulfilment of the requirements for the degree of

Master of Science in the Faculty of Science

at the

Stellenbosch University

Supervisor: Dr Cang Hui

April 2014

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Marinel Janse van Rensburg
Stellenbosch, April 2014

ABSTRACT

Much emphasis has been placed on modelling species distributions and less so on the distributional dynamics. It is important to not only ask where species occur, but why and how they reached a specific location. Biological invasions provide an ideal natural experiment for studying the drivers and dynamics of spreading. Here, I examine the drivers and dynamics of the spread of the European Starling, *Sturnus vulgaris*, in Southern Africa. Since its introduction in Cape Town in 1897, this top invasive avian species has rapidly spread across a large extent of South Africa and is continuously expanding its current range. It is, thus, of great theoretical and management value to elucidate the invasion process and identify key environmental drivers of its range dynamics.

The aim of this research is twofold. First, I aim to develop realistic suitable habitat maps for the European Starling in Southern Africa based on multiple climatic and geographic variables using two robust methods of species distribution modelling (SDM), namely maximum entropy (MaxEnt) and boosted regression trees (BRT). Second, I aim to, subsequently, develop a hybrid model that runs a dynamic individual-based model (IBM) on these suitability maps predicted from the SDM. This hybrid model provides an excellent opportunity to incorporate a variety of response regimes on how behavioural strategies and dispersal scenarios respond to environmental and geographic features, and as a result leads to a well-rounded study of the species' distribution dynamics in the region. Independent atlas data and field collections were used to parameterize and later validate the model. This allowed me to further identify an optimal model through a parameter sensitivity analysis sorting of the contribution of each environmental and behavioural features to shaping the past and current geographical range of European Starling.

According to the optimal model, starlings can choose among five sites to locate the one with the highest habitat suitability. They tend to avoid moving through areas with a 300 m elevation barrier, and the dispersal distance per year is limited to below 200 km. At the regional scale, the starlings were able to manoeuvre around mountainous regions and avoided the semi-desert regions of the Karoo. Their distribution was mainly driven by high winter precipitation along the low-lying coastal regions. Future projections of their distribution suggested a continuous range expansion throughout the provinces of the Free State, Gauteng, North West (reaching Mahikeng, the capital of North West, in 2022) and Limpopo (reaching Polokwane, the capital of Limpopo, in 2046) as well as into the neighbouring countries of Mozambique and Zimbabwe (reaching Bulawayo, 2nd largest city in Zimbabwe, in 2062).

OPSOMMING

Klem word dikwels meer geplaas op die modellering van spesies se verspreiding en minder op die dinamika agter die verspreiding. Dit is belangrik om nie slegs te vra waar spesies voorkom nie, maar hoekom en hoe hulle by 'n spesifieke plek beland. Biologiese indringing bied 'n ideale natuurlike eksperiment vir die bestudering van die bestuurders en dinamika agter spesies se verspreiding. Hier ondersoek ek die bestuurders en dinamika agter die verspreiding van die Europese Spreeu, *Sturnus vulgaris*, in Suider-Afrika. Sedert sy bekendstelling in Kaapstad in 1897, het hierdie top indringer voël vinnig versprei oor 'n groot gebied van Suid-Afrika en is tans steeds besig om verder te versprei. Dit is dus van groot teoretiese en beskermings waarde om die indringers proses te verstaan en die noodsaaklikste omgewings bestuurders van die spesies se verspreidings dinamika te identifiseer.

Die doel van hierdie navorsing is tweeledig. Eerstens, streef ek daarna om realistiese geskikte habitat kaarte vir die Europese Spreeu in Suider-Afrika te ontwikkel wat gebaseer is op verskeie klimaat en geografiese veranderlikes met behulp van twee robuuste metodes van verspreiding spesies modellering (*species distribution modelling*, SDM), naamlik maksimum entropie (MaxEnt) en regressie bome (*boosted regression trees*, BRT). Tweedens, streef ek daarna om 'n hibriede model te ontwikkel wat 'n dinamiese individu-gebaseerde model (IBM) op hierdie geskiktheids kaarte, voorspel deur die SDM, te simuleer. Die model bied 'n uitstekende geleentheid om 'n verskeidenheid van reaksie kombinasies te integreer oor hoe gedrags-strategieë en verspreiding scenario's reageer op omgewing en geografiese kenmerke, en as 'n gevolg lei tot 'n afgeronde studie van die spesie se verspreiding dinamika in die streek. Onafhanklike atlas data en veld data is gebruik om die model te parameteriseer en die geldigheid daarvan te toets. Hierdeur kon ons ‘n optimale model identifiseer deur middel van 'n parameter sensitiwiteits-analise sortering van die bydrae van elke omgewings en gedrags kenmerk tot die vorming van die verlede en die huidige geografiese verspreiding van die Europese Spreeu .

Volgens die optimale model, kon die spreeus tussen vyf plekke die een met die hoogste habitat geskiktheid kies. Hulle is geneig om gebiede met 'n 300 m hoogte versperring te vermy, en die verspreidings afstand per jaar is beperk tot minder as 200 km. Op die plaaslike skaal, die spreeus was in staat daartoe om te beweeg rondom bergagtige gebiede en die semi-woestyn gebiede van die Karoo te vermy. Hul verspreiding is hoofsaaklik gedryf deur 'n hoë winter reënval langs die laagliggende kusgebiede. Toekomstige projeksies van hul verspreiding stel voor dat verdere verspreiding plaasvind in die provinsies van die Vrystaat, Gauteng, Noord-Wes (waar hulle Mahikeng, die hoofstad van Noord-Wes, bereik in 2022) en Limpopo (Polokwane, die hoofstad van Limpopo, word bereik in 2046) sowel as in die buurt lande van Mosambiek en Zimbabwe (Bulawayo, 2 grootste stad in Zimbabwe, word bereik in 2062).

ACKNOWLEDGEMENTS

It is with immense gratitude that I acknowledge the support and advice of my supervisor, Dr Cang Hui, without whom this thesis would not have been such a success. I am indebted to my colleagues and fellow members of the Theoretical Ecology Journal Club who provided many hours of thought-provoking discussions.

This thesis would not have been possible without the financial support from the Council for Scientific and Industrial Research, CSIR.

I am particularly grateful for the Centre of Excellence for Invasion Biology for their support and the travel bursary awarded to me. This bursary funded my travels to the University of Alberta, Canada, where I was able to attend a summer school on the Mathematics behind Biological Invasions. I would hereby also wish to thank the organizers and sponsors of the summer school: Pacific Institute for the Mathematical Sciences, Centre for Mathematical Biology, University of Alberta and Mathematics of Planet Earth 2013.

I would like to acknowledge: Prof. Mark Robertson for his assistance with the MaxEnt software and for organising the workshop on Species Distribution Modelling at the University of Pretoria; the Centre for Geographical Analysis for presenting an in depth Introductory Course on ArcGIS; the Animal Demography Unit for providing the SABAP data; Dr Cecile Berthouly-Salazar for her advice and their research group for independent data collection.

I wish to express my deepest appreciation to the endless support and love received from my family, friends, Abraham and his family. They made this journey worthwhile and their continuous motivation provided the backbone for this end product. Special thanks are extended to Abraham for his encouragement and multiple editing sessions.

TABLE OF CONTENTS

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADU	Animal Demography Unit
ANN	Artificial Neural Networks
AUC	Area under the Receiver Operating Curve
BRT	Boosted Regression Trees
CA	Cellular Automaton
CGIAR-CSI	Consortium for Spatial Information of the Consultative Group for International Agricultural Research
CIESIN	Centre for International Earth Science Information Network
CV	Cross-Validation
DEM	Digital Elevation Map
GAM	Generalized Additive Models
GARP	Genetic Algorithm for Ruleset Prediction
GIS	Geographic Information System
GLC	Global Land Cover
GLM	Generalized Linear Models
GME	Geospatial Modelling Environment
HFP	Human Footprint
HP	Hierarchical Partitioning
HS	Habitat Suitability
IBM	Individual Based Model
IPS	Interacting Particle System
MARS	Multivariate Adaptive Regression Splines
MaxEnt	Maximum Entropy
MESS	Multivariate Environmental Similarity Surfaces
OLS	Ordinary Least Squares Regression
RF	Random Forests
RMA	Reduced Major Axis Regression

ROC	Receiver Operating Curve
SABAP	South African Bird Atlas Project
SANBI	South African National Biodiversity Institute
SDM	Species Distribution Model
TSS	True Skills Statistic
VIF	Variance Inflation Factor

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

A lot of emphasis is often placed on modelling species' distributions, and less on their distributional dynamics. It is important to not only ask where species occur, but why they occur at a specific location and how they got there. Scientists are striving to answer these questions and understand the patterns that are observed during species' range expansions. Another ecological phenomenon that causes a variety of questions and concerns is invasive species. The impacts of invasive species on their invaded surroundings are enormous and continue to escalate. The growing degree to which the whole planet is connected allows for easy transportation and spread of species between locations, subsequently increasing the number of introduced and potentially invasive species. Biological invasions provide an ideal natural experiment for studying the drivers, mechanisms and dynamics of spreading, and predicting the spatial spread of invasive species is crucial for management purposes. Due to technological advances we are able to develop models that enable us to understand and predict these processes.

Models used for studying species' range expansions were typically based on reaction-diffusion or integro-difference equations (Hui *et al*., 2011). They lacked however the practicability of incorporating complex spatial components such as heterogeneous environments which would enable them to resemble more closely the aspects of real landscapes. Incorporating spatial components along with demographic components are crucial for studying the invasion process if one aims to closely predict the spread of invasive species. Modelling methods consequently developed in a direction away from general equations towards models that could represent each individual or group of individuals separately such that the dynamics of the greater system of populations are captured as a whole (Judson, 1994). More recently, the modelling focus has moved towards cellular automata (CA) and individual based models (IBM), also collectively known as interacting particle systems (IPS). In an IPS the landscape is presented as grids on a two-dimensional space. The main drivers behind the spread of the species are their dispersal kernel and demographic rates. Within these frameworks one could even add another spatial dimension that takes into account the habitat suitability of the species being studied, one way of approaching this is through niche modelling.

Niche modelling has opened up new doors for studying population distributions. There are two main approaches for predicting species' niches (Gallien *et al*., 2010). Firstly there's the bottom up approach (mechanistic), which uses the physiological characteristics of a species to determine their suitable habitat. Secondly there's the top down approach (correlative), which focuses on the species-environment relationship and the associations between the species' distribution and the environmental factors. Climate is often modelled as the main driver behind species' distributions, but their distributions are in actual fact co-determined by climate, physical structures, disturbances and biotic and abiotic interactions. When modelling species' distributions mechanistically, climate data derived from weather stations needs to be translated into microclimate data as organisms respond morphologically and physiologically more to

microclimates than to climate conditions as measured by weather stations (Kearney and Porter, 2009). Mechanistic niche modelling therefore gets highly complicated to simulate due to this challenging data collection procedure.

Niche models on their own are not highly efficient for studying species' range expansions, but when combined with an IPS considerably more can be learnt about the distribution dynamics of the species (Carey, 1996). The result of this is a hybrid modelling framework which closely resembles both the species' and the environment's characteristics, ensuring a more realistic model. With such a dynamic model other features can be incorporated, for example dispersal kernels, behavioural rules, geographic structures and habitat suitability, and one can study how they influence the predictions. Individuals can be tracked and the rate at which species will spread or decline can be assessed.

Mechanisms behind the range expansion of invasive species are related to dispersal strategies such as fat-tailed dispersal, spatial sorting or flexible dispersal strategies triggered by habitat quality and surroundings. Different phases of the invasion process are associated with different distribution dynamics. The initial establishment phase is determined by diffusion of the founder population, which is followed by an accelerated expansion phase determined by population growth and long distance dispersal events (Shigesada *et al*., 1995). The nature of the expansion process depends on the life-history characteristics of each species. Throughout the invasion process various dynamics take place. Species often undergo a niche change, where new environmental and geographic landscapes are encountered as well as new biotic interactions. All these factors lead to a different system than the one from their native habitat. Certain species are able to adapt to these changes, others not. Combining all of these processes leads to a study of distribution dynamics of the invasion process. Limited modelling exists regarding the study of invasive species with dynamic models, and less so where mobile species are concerned (Higgins *et al*., 1996a; Jongejans *et al*., 2008; Nehrbass and Winkler, 2007; Travis *et al*., 2007).

1.2 Approaching the Problem

According to Hui *et al.* (2011) an optimal model would be one that:

- describes the demographic and spatial dynamics of the study species,
- can identify ecological determinants and mechanisms of the spatial-temporal dynamics,
- investigates factors affecting the rate of spread and spatial distribution, and
- studies biotic interactions, environmental heterogeneity and stochasticity with the goal of designing spatially optimal strategies for detection, control and eradication of invasive species.

Following this modelling framework, I illustrate in Figure 1.1 my optimal model for studying the spread of an invasive species. Species distribution models (SDM) provide the modelling environment where important predictor variables are investigated for the species' distribution and a suitable habitat map is obtained. This suitable habitat output from the SDM is employed as an input into the dynamic individual based model which serves as the modelling core. Within this core the dynamic model further includes spatially explicit factors such as geographic barriers, behavioural rules like cognitive ability, demographic processes and dispersal strategies. The spatial-temporal facet of the model allows me to track individuals through time in order to study their dispersal routes. Keeping track of the range front and how it changes through the invasion process allows me to analyse the different phases associated with the range expansion. Using an IBM to model species' distributions is a very efficient procedure for visualising the species' range expansion as well as obtaining insight into the leading forces behind their distribution.

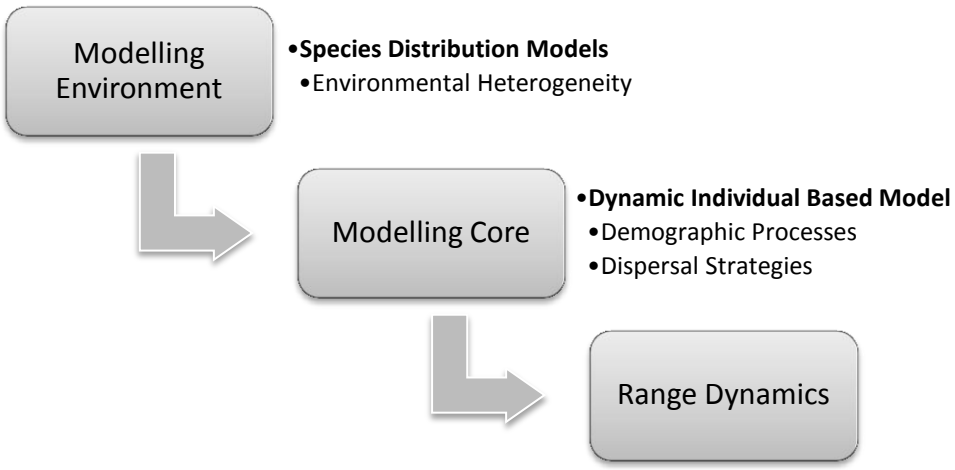


Figure 1.1: A framework for modelling the range expansion of an invasive species which was applied in this study.

The European Starlings' spread in Southern Africa can be used to examine the above-mentioned distributional dynamics of an invasive species. Since its introduction in the Western Cape province in 1897, the European Starling, *Sturnus vulgaris*, has rapidly spread across a large extent of South Africa and is continuously expanding its current range. It is of great theoretical and management value to elucidate its invasion process and identify key environmental drivers of the range dynamics. Knowledge is limited regarding the starling's distributional dynamics in Southern Africa and a hybrid modelling approach that incorporates their dispersal strategies with environmental and geographical structures is lacking. This dynamic hybrid method allowed me to reconstruct the history of the starling's range expansion in Southern Africa during the last century and provided a way to forecast future range dynamics.

My objectives were defined as follow:

Objective I: *Construct a potential distribution for the European Starling in Southern Africa using a species distribution modelling approach.*

To begin with, an in depth study of the European Starling was necessary and this is discussed in Chapter 2 (*Sturnus vulgaris*) along with the species' datasets that were implemented in the modelling approach. Applying this knowledge, appropriate distribution models were developed and are shown in Chapter 3 (Species Distribution Models).

Objective II: *Reconstruct the European Starling's range expansion using a dynamical modelling approach.*

The potential distributions obtained in Chapter 3 were used in the construction of the dynamic individual based model in Chapter 4 (Individual Based Models). This model was used to study the distribution dynamics of the European Starling and parameter sensitivity analyses were carried out to obtain models that best predicted the observed distribution of the starling in Southern Africa.

Objective III: *Forecast future distributions of the European Starling in Southern Africa.*

The top performing models were used to forecast future range expansions of the European Starlings and to study their dispersal routes, also discussed in Chapter 4.

CHAPTER 2

STURNUS VULGARIS

2.1 Characteristics

The European Starling, *Sturnus vulgaris*, belongs to the family Sturnidae, order Passeriformes. Adult starlings have a body length, from head to tip of tail, of approximately 21.5 cm with an average wingspan of 40 cm and weigh 70 to 100 grams (Chow, 2000; Lynch and Messmer, 2010). Males and females look very similar, they have dark glossy feathers with yellow and brown shaded flecks and a short tail (Figure 2.1). Juveniles have brownish beaks whereas adults have a yellow beak during the breeding season which turns dark brown during the rest of the year. Females often choose their mates based on the male's ability to defend his resources and the quality of resources he has available. They are highly vocal using sounds varying from a click, warble, creak or chirrup. They are also great imitators, mimicking sounds from sirens to croaking frogs.

Figure 2.1: The European Starling in its introduced habitat, South Africa. Photos were taken in Sea Point, Western Cape province.

2.1.1 Habitat and Feeding

European Starlings are habitat and dietary generalists, which is one reason why they are such successful invaders. They are often found in areas where humans are present, while they tend to avoid deserts and mountainous areas (Link, 2004). They are one of the few bird species that can tolerate highly populated areas and poorly vegetated landscapes such as industrial sites. Starlings are excellent flyers, reaching

speeds of up to 60 - 80 km/h, and migrate distances up to 1500 km to find suitable food sources and nesting sites. Migration patterns vary by year and location, but they rarely migrate south of 40˚N (Linz *et al*., 2007). In North America their migration may take place either during Spring, February to March, or during Fall, September to December (Kessel, 1953).

Starlings commonly forage on short grass areas such as lawns and pastures where they eat plant matter and invertebrates, but they are also fruit and seed eaters (Link, 2004; Linz *et al*., 2007). They are attracted to cultivated lands because irrigation allows for easier probing for invertebrates in the moist soil (Harrison *et al*., 1997). They can form feeding and roosting flocks of hundreds to thousands of starlings and during winter these roosts can become as large as one million starlings (Lynch and Messmer, 2010). When forming such large groups they can perform spectacular aerial displays of coordinated movements. They form such roosts to exploit abundant food resources and once established at a feeding site it is difficult to scare them away despite tremendous efforts from humans (Linz *et al*., 2007).

Fischl and Caccamise (1987) found that in North America the starlings' diet by dry weight consist on average of 62% plant matter and 21% animal matter, which can vary between different seasons. These values do not necessarily indicate that their diet consists mostly of plants as undigested plant material comprises greater mass than undigested animal material. Early in the roosting season when flock sizes are still small their diet contains more animal matter than later in the roosting season. The reason for this is that with larger flocks they possess greater ability to search for and exploit suitable lands for feeding.

2.1.2 Nesting and Reproduction

European Starlings nest in holes or crevices in buildings, structures, trees or even cliffs (Link, 2004; Linz *et al*., 2007). They are known to be secondary cavity nesters, which mean they do not make their own holes and nests, but take control of other birds' nests. The male starlings can be very aggressive when claiming another bird's nesting spot as their own. In North America starlings will commonly form large flocks and roost in either trees or barns, depending on the season (Link, 2004; Lynch and Messmer, 2010).

A study on the breeding behaviour of starlings in New York found that individuals were very likely to remain close to their nesting site once they had raised a brood there, with 30% of females using the same nesting box in consecutive years while 90% moved less than 1 km to breed (Kessel, 1957). It was found that the majority of dispersal events were natal dispersal where the juveniles would disperse far and wide in search of new breeding sites.

In North America starlings lay eggs anytime from February to July (Link, 2004; Linz *et al*., 2007). A clutch size is commonly between four to six eggs which hatch after an incubation period of eleven to

thirteen days. The female does all of the incubation during the night and about 70% of the incubation during the day. The young will leave the nest twenty to thirty days after hatching, after which the mother sometimes lay a second clutch of eggs. In Sweden they breed exclusively in May and rarely have second breeding seasons in the same year (Lynch and Messmer, 2010). The starling's average lifespan is two to three years, but there have been individuals recorded to live as long as twenty to thirty years in Australia (Linz *et al*., 2007; Waterman *et al*., 2008).

Adult starlings have very few natural predators, another attribute which allows them to invade with high success. Hawks and falcons have occasionally been seen to catch them in flight (Link, 2004). The biggest threat to starlings is humans, who try to manage them via control programs. Other possible limitations to their population sizes could be the availability of nest sites, extreme weather events that limit their food sources for example temperature sensitive invertebrates, and squirrels accessing their nests and destroying clutches (Linz *et al*., 2007). Kessel (1957) found that starlings in North America had a 48% to 79% nest success rate, while only 20% of nestlings survived to reproduce. Adult survival was found to be much higher at around 60% (Linz *et al*., 2007).

2.2 Native Habitat and Introduced Regions

The European Starling is one of three birds listed on the World's Worst Invasive Alien Species list (Lowe *et al*., 2000). They are native to Europe and western Asia, migrating to North Africa and the Middle East during winter. Apart from this natural habitat, European Starlings have been introduced to North America, New Zealand, Australia and South Africa, where approximately 80% of introduced populations have become established (Chow, 2000; Long and Tingay, 1981; Sol *et al*., 2002) (Figure 2.2). Some of these introduced populations may migrate over short distances, but in warmer climates such as South Africa, they are non-migratory.

In North America the starlings' range expanded much more rapidly than in South Africa. About 60 starlings were first released in New York's Central Park in 1890 by Eugene Schlieffelin (Link, 2004; Schuster, 2010). The reason behind their release was to introduce all the species mentioned in Shakespeare's plays to the USA. Since their introduction the starlings have spread up north into Canada and Alaska and southward into Central America. Today, it is estimated that there are about 200 million starlings in North America (Linz *et al*., 2007).

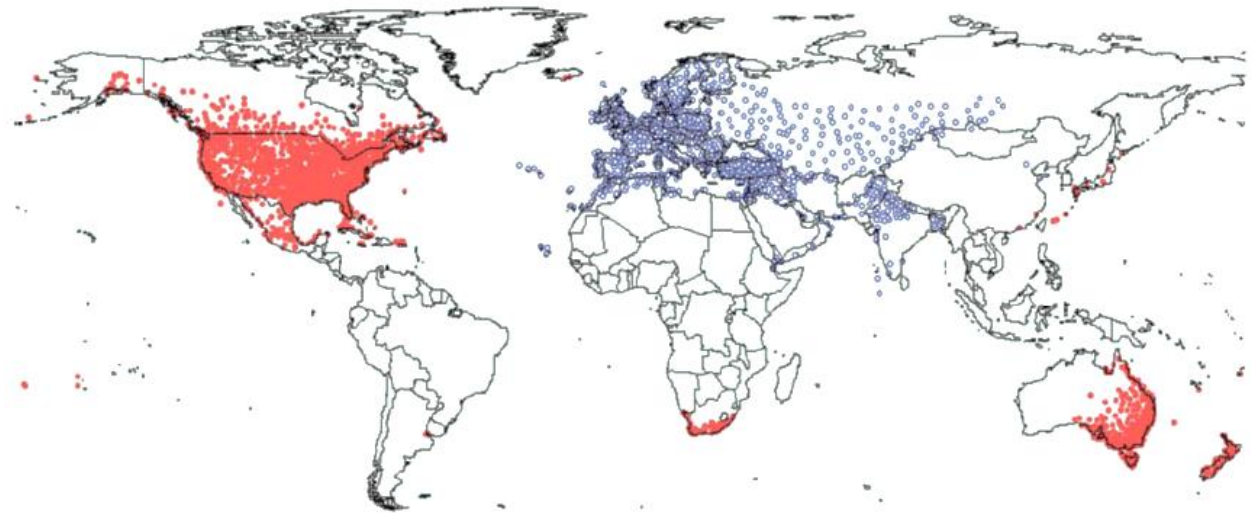


Figure 2.2: Worldwide distribution of the European Starling with native ranges in blue and introduced ranges in red (Kirkpatrick and Woolnough, 2007)**.**

2.2.1 Invading South Africa

Cecile John Rhodes introduced the European Starling to South Africa in 1897 when releasing 18 individuals in Cape Town (Sinclair and Davidson, 2006). Since then the starlings have spread across South Africa and neighbouring Namibia and their range still continues to expand. Initially their range expanded slowly outwards from the Cape Peninsula and was recorded in Stellenbosch in 1908 and Somerset West in 1910. They spread much faster along the eastern coastlines than the western coastlines, reaching both Groot Brakrivier (406 km to the east of Cape Town) and Kogmanskloof (179 km to the north of Cape Town) in 1940. Expanding still further eastwards they reached Port Elizabeth in 1954, East London in 1966 and Maclear in 1982 (Figure 2.3). Overall their spread has been much slower than in North America. The starlings were introduced in North America only seven years earlier than in South Africa and in Figure 2.2 we can already see the vast difference in their current distribution. This could be due to the initial number of birds released (18 in South Africa versus 60 in North America), or environmental and geographical factors that were more accommodating in North America. The European Starling is non-migratory in South Africa due to the warm climate, however, they do still form flocks during winter time (Craig, 1997). Breeding season is September to December.

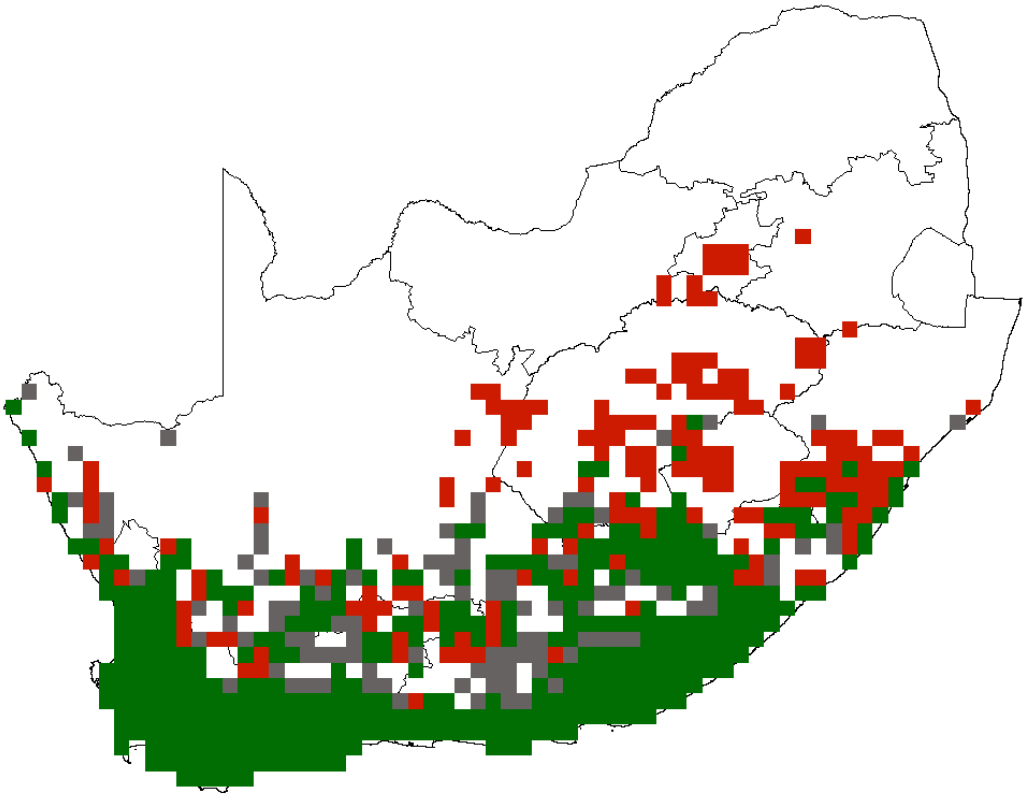


Figure 2.4: Comparing SABAP 1 (1997) and SABAP 2 (2012) presence records for the European Starling at a 15 minute degree resolution. Green grids are where SABAP 1 and SABAP 2 overlap, grey grids are only SABAP 1 records and red grids only SABAP 2 records.

The majority of starling occurrences in Southern Africa are associated with grasslands, shrublands, croplands and deciduous woodland, while deserts and submontane forests are avoided according to land cover data (GLC2000, 2003), SABAP 1 and SABAP 2 (Figure 2.5). This supports Link's observation (2004) that starlings are generally absent from deserts and heavily wooded areas. They are however present in deciduous woodland and closed deciduous forests in South Africa, suggesting that a possible niche shift took place. Figure 2.6 illustrates the relationship between the relative density of the starling and elevation, distinctly showing their preference for low lying areas. All of the life-history parameters used in this study regarding the European Starling in South Africa were obtained from Hui *et al.* (2012) and are shown in Table 2.1.

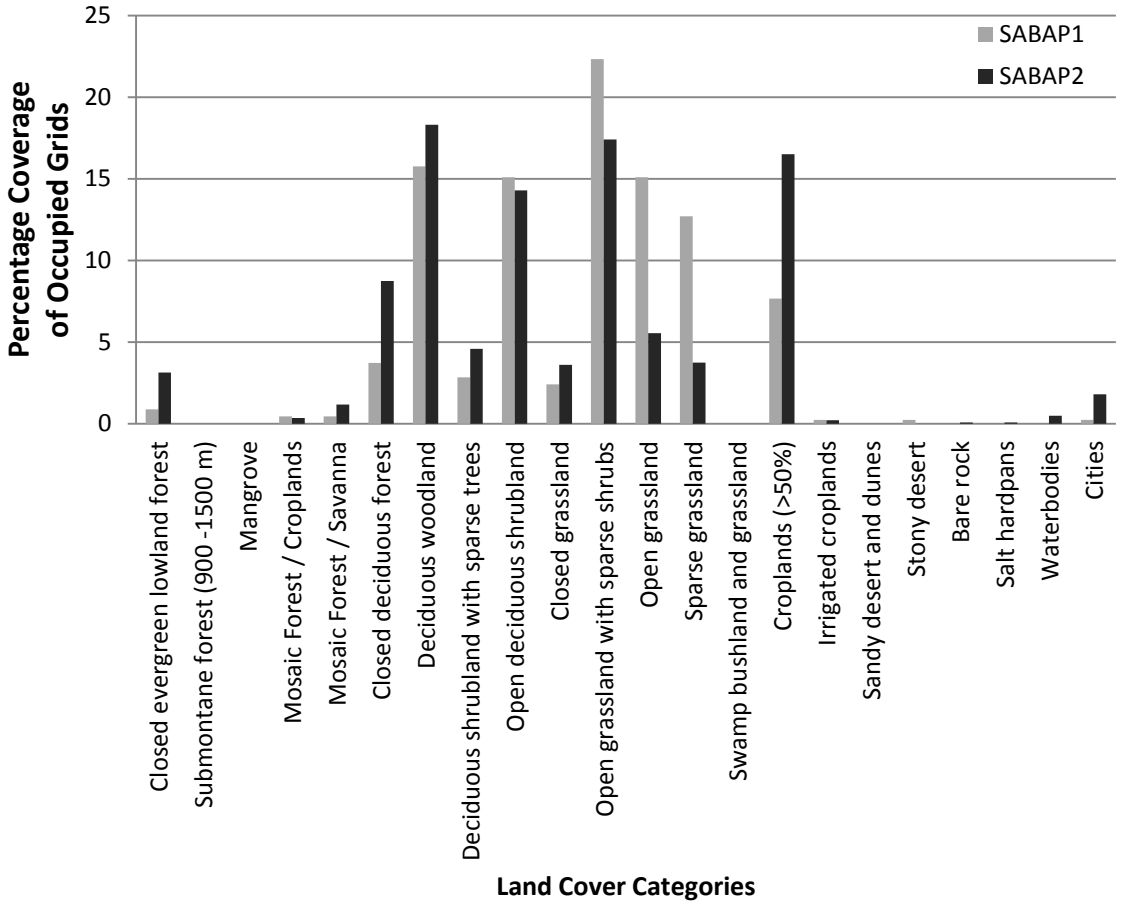


Figure 2.5: The percentage coverage of grids occupied by the European Starling from both SABAP 1 and SABAP 2 according to different land cover categories.

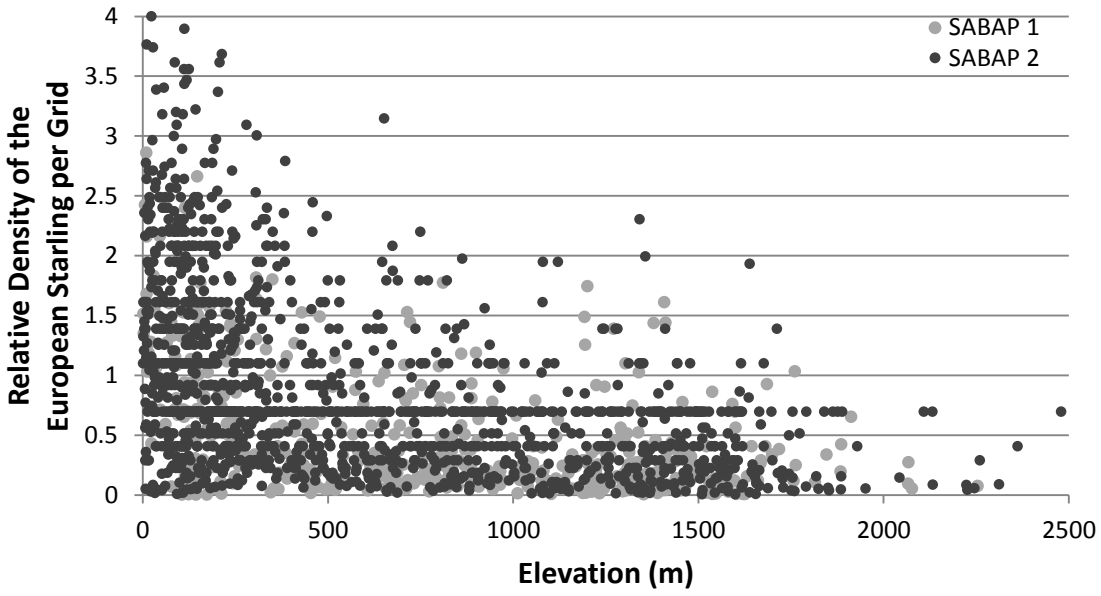


Figure 2.6: Relationship between the European Starling's relative density and elevation for SABAP 1 and SABAP 2 presence records.

Table 2.1: Life-history parameters of the European Starling in South Africa (Hui *et al*., 2012)**.**

Intrinsic growth rate per year	(0.110, 0.130)
Adult survival rate per year	0.647
Juvenile survival rate per year	0.324
Clutch size	4.400
Nest success	0.740

2.2.2 Problems of Invasion

One of the introduced ranges that experience a vast array of problems from the establishment of this invasive species is North America (Ghianni, 2013). Pimentel *et al.* (Pimentel *et al*., 2000) estimated the yearly damage to agricultural lands by starlings to be in the order of US$800 million. They attack both fruit and grain crops. Not only do the starlings have a tremendous impact on the economy, they also spread diseases and compete aggressively with native birds for nest cavities (Ghianni, 2013). In Australia the latter is particularly problematic and it was found that the decline of certain native bird species such as parrots is due to European Starlings (Kirkpatrick and Woolnough, 2007). In South Africa, no evidence has yet suggested that the decline of native bird species is due to competition with starlings. Most declines in native species are believed to be due to habitat change.

Changing climate is sending starlings to places where they are not commonly seen, bringing along with them an assortment of problems and diseases. One example is Hopkinsville in Kentucky, USA, which is used to experiencing winters with temperatures below 0 degrees Celsius (Ghianni, 2013). In February 2013 it was not as cold as usual, and where the starlings and blackbirds would normally go further south for the winter, they remained in Hopkinsville in their large numbers. They became a pest, adding to noise and ground pollution. The city organised pest controllers to take charge, but their techniques of blasting cannons to scare away the birds were scaring other animals away along with the starlings. The amount of excrement they left behind posed big health risks and the city, among others, struggled to get rid of the large flocks of birds.

Starlings can transmit diseases to livestock as well as humans and other animals (Lynch and Messmer, 2010). Combined faecal matter can release fungal spores. When these spores are airborne and inhaled by humans it can cause histoplasmosis, which is a respiratory disease that has on very rare occasions even caused blindness and death in humans. To date there has been eight bacterial diseases, six fungal diseases, four protozoan diseases and six viral diseases which could be transmitted from starlings to humans and animals (Linz *et al*., 2007; Lynch and Messmer, 2010). Most transmission occurs through inhaling these pathogens that live in dried faeces. Johne's disease in cattle, where the bacteria is excreted

in faeces and milk, caused the US dairy industry $200-250 million annually. Furthermore, the acidity of the starling droppings can corrode metals causing damage to infrastructure.

The starlings have been observed to consume up to 15 to 20 tons of cattle feed per day when they form large roosts (Linz *et al*., 2007). This can accumulate into vast economic losses. Starlings that feed on fruit crops may damage the quality of the produce and impact yields. Studies have shown that in the US in 1972, the loss of grapes by birds accumulated to $4.4 million and 17% of cherry crops were attacked by starlings.

As with other birds, starlings can pose a safety hazard to aircrafts if caught in the jet engines and this problem intensifies in areas where starlings form enormous roosts and dense flocks. Starlings were the cause of an aircraft to crash in Boston in 1960 which resulted in the deaths of 62 people (Linz *et al*., 2007).

2.2.3 Managing the Invasion

Since the European Starling is such a big problem in North America, a vast amount of notices and advice has gone out to citizens on how to prevent starlings from establishing a foothold in their area, including: how to close up certain nesting sites or make nesting holes small enough only allowing small birds access, how to build feeders that will exclude starlings, how to protect native cavity nesters from being outcompeted by starlings, how to protect fruit crops and how to trap, repel or kill starlings (Link, 2004; Lynch and Messmer, 2010).

Pyro techniques, propane exploders, kites and ultrasonic sounds are some common deterrents used to physically frighten the starlings (Linz *et al*., 2007). However, most of these techniques are only temporary and rarely gets rid of the starlings. It was found that the starlings would just move from one spot to the next and were likely to even return to the same spot after the deterrents had worn off. Chemical agents such as pesticides are also used to frighten them. One such agent is 4-aminopyridine, which is put in grain or pellets and could result in the death of the bird digesting it. This is risky since other non-targeted birds could also die from these treatments as well as animals eating the poisoned birds. It is worrying to admit, but at the same time it is the reality, that “there is no prospect of eliminating this resourceful bird” (Craig, 1997).

2.2.4 Dispersal Strategies

The reproductive biology regarding the European Starling has been well researched, but what is lacking and what is needed in order to assist management strategies in controlling the starlings and developing risk assessments is knowing their migratory and movement patterns (Linz *et al*., 2007). This is where physical tracking methods are required. However, due to the difficulty of implementing such a cost and labour intensive method, computational modelling can be used as a substituted tracking method.

Cabe (1999) studied the dispersal of the European Starling in North America using both band return analysis (tracking the birds) as well as genetic analysis and found that natal dispersal of juveniles ranged anything between 12 – 2 623 km with an average of 219 km. The distribution is both skewed and highly leptokurtic. Adult starlings rarely changed breeding sites once they had bred in a specific area while the juveniles were the ones found to travel further away. In Australia, Waterman *et al.* (2008) banded and recaptured starlings and found that some individuals travelled 60 - 90 km within one to two months. One individual was captured 950 km away from the banding site, only four months later.

The European Starling's behaviour was recorded in a Swedish study and it was found that certain individuals in the population acted as floaters (Tobler and Smith, 2004). Floaters explore other sites to find information about possible nesting sites and start acquiring a local dominance position, thereby giving them the added advantage of having access to certain nest sites in the future. For secondary nesters such as starlings, it is important to acquire this information about other areas in terms of the number of available nest sites and suitability for reproduction.

The most commonly used dispersal kernels to describe animal movement are the negative exponential function, $f(d) \sim e^{-ad}$, and the inverse power function, $f(d) \sim d^{-a}$, where f is a function of d, distance, and a is a parameter that determines the width of the dispersal kernel's tail. Hui *et al.* (2012) used ringing records and atlas data to establish that the starling's dispersal kernel resembled an inverse power law function, $f(d) \sim d^{-1.5}$. An exponent less than 2, as in this case, indicates a fat-tailed, long-distance dispersal kernel. The native dispersal kernel was found to be significantly steeper than the non-native kernel. This difference suggests a flexible dispersal strategy, where the starlings respond to spatial and temporal variation and adjust their strategies accordingly. Hui *et al.* (2012) suggested a 'good-stay, bad-disperse' strategy in which dispersal rates increase following a decrease in environmental quality. Furthermore, introduced species often experience a niche shift between their native and introduced regions, therefore we would expect a shift in their dispersal strategies. Such shifts could be either changing dispersal distances or changing the number of long distance dispersal events. The wave speed of an invasion is mostly determined by the upper percentiles of the dispersal distribution, hence long distance dispersal events. Dispersal strategies have previously been identified as playing a more important role than demographic factors in the range dynamics of species' expansion, including those of the European Starling (Caswell *et a*l., 2003; van den Bosch *et al*., 1992).

2.3 Data Collection and Processing

2.3.1 SABAP 1

The First Southern African Bird Atlas Project (SABAP 1) launched in 1987 and collected bird distribution data throughout six Southern African countries over a five year period, continuing up to 1991 (Harrison *et al.*, 2008). This project was initiated by a combined effort of the Animal Demography Unit (ADU) at the University of Cape Town and the South African National Biodiversity Institute (SANBI). The countries it covered were South Africa, Lesotho, Swaziland, Namibia, Botswana and Zimbabwe. Unfortunately Mozambique could not be included due to the civil war at that time. The advantage of data collection taking place over several years was that seasonal variation could be studied in different bird species, accounting for breeding seasons and migrations.

The end result of this project was a publication in 1997, *The Atlas of Southern African Birds*, with an accumulation of seven million bird distribution records (Harrison *et al.*, 2008, 1997). This data was collected by 'citizen scientists', the term following the concept that any citizen can be a scientist by participating and helping collect data in a rather straightforward process (described in the following paragraph). After this bird atlas, others followed, including frogs, proteas, reptiles, spiders and butterflies. SABAP was the first biodiversity study of its scale to be accomplished in Africa and received several awards of which the most notable was the John FW Herschell Gold Medal of the Royal Society of South Africa given to the authors in 1999 (ADU). Due to the need for such data and its application in different fields, from research to policy making, online publication of the atlas and certain data sets have been made available to the public.

The SABAP 1 data was collected at a 15 minute grid resolution (about 25 km x 25 km in South Africa). Each region had its own checklist containing a list of species that were common to that region which volunteers used for collecting the data. Before commencing this process, each volunteer received an instruction booklet on the necessary steps to follow. Observers were encouraged to submit an unbiased recording list of all species observed in a grid in order to obtain lists that were representative of the grids. Additional information could be added such as whether the bird was only seen or heard, whether the observer suspected breeding was taking place, eggs were present, chicks were present, etc. The recording sheets were not automatically entered into the atlas database as a large degree of quality control took place first (Harrison *et al.*, 1997). Unusual records were evaluated with additional information provided, and scrutinized in greater detail before being accepted into the database.

There were many remote areas that remained unexplored. Most volunteers completed recordings within their areas of residence or easily accessible regions. To improve the recording across all regions, various strategies were employed such as sponsoring petrol expenses for remote areas, organizing group expeditions and employing fieldworkers (Harrison *et al.*, 1997). The coverage and distribution of the

whole region used for the SABAP 1 collection period is shown in Figure 2.7. Those records that sighted the European Starling are shown in Figure 2.8.

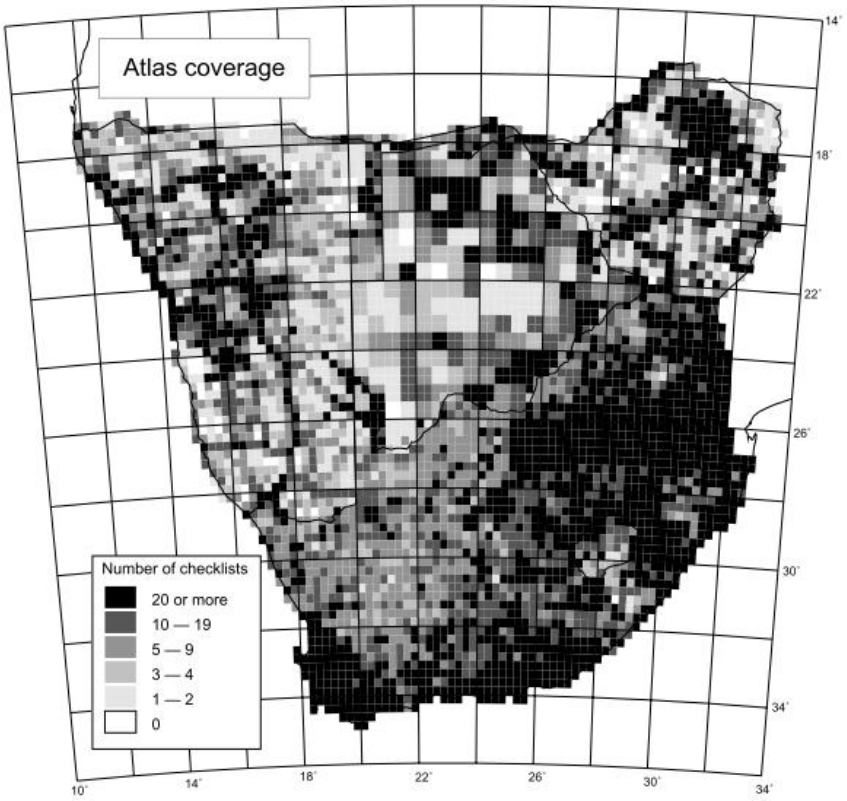


Figure 2.7: Number of checklists per grid cell after completion of SABAP 1 in 1991 (ADU).

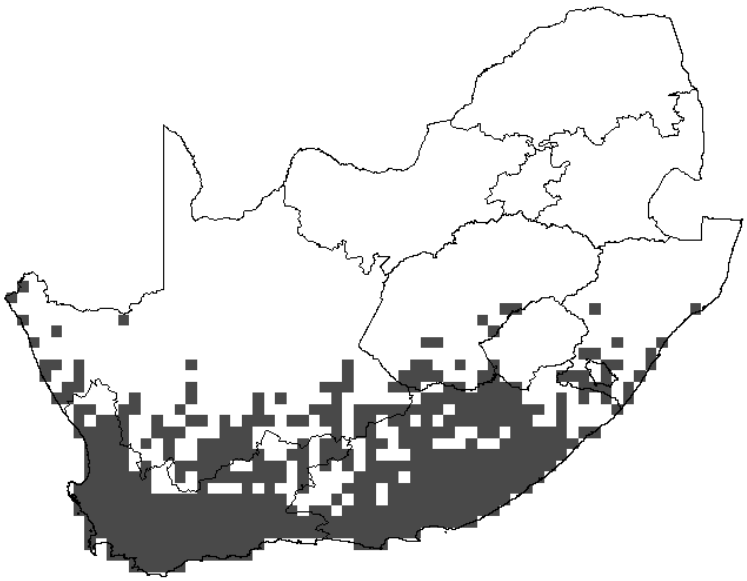


Figure 2.8: Presence records of the European Starling in South Africa in 1997 as collected by SABAP 1 at 15 minute degree resolution.

2.3.2 SABAP 2

Due to SABAP 1's success and the necessity to accumulate even more data and study changes in bird populations over time and space, the Second Southern African Bird Atlas Project (SABAP 2) commenced in July 2007. It is a partnership project between SANBI, ADU and BirdLife South Africa, but unlike SABAP 1 it only covers South Africa, Lesotho and Swaziland (ADU). As with SABAP 1, SABAP 2 also makes use of citizen scientists to collect the data and thereby raising awareness and increasing public's interest in birds. It is an on-going project and since its beginning up to November 2013 there have been a total of 5.04 million records covering 71.78% of the total pentads in the study region. Regular atlas workshops and presentations are held across the country for volunteers to attend in order to ensure better understanding of the atlasing procedure and improving their data collection skills.

SABAP 2 improved on SABAP 1 by collecting data on smaller grids. Instead of the 15 minute degree resolution used in SABAP 1, SABAP 2 uses grids of 5 minute degree resolution (about 8 km x 7.6 km in South Africa). Furthermore, the data collection has been improved by completing an in depth field study while still keeping the process straightforward. Observers now need to record for at least two hours and for every species recorded they need to state in which hour they were recorded. Observers also have to identify the type of habitats they are visiting and any additional observations. All of the records are submitted electronically, processed immediately and loaded onto the ADU's SABAP 2 website which continuously keeps all of the statistics and projections up to date. As with SABAP 1 it also undergoes a quality control procedure where the data are first assessed and verified. Figure 2.9 shows the coverage of SABAP 2 from 2007 to November 2013 in the data collection procedure. The distribution of the European Starling according to this data is shown in Figure 2.10.

The number of individuals per unit area, referred to as density, is complicated to estimate. I therefore estimated the starlings' relative density based on the Poisson model. The Poisson distribution assumes that the individuals are randomly distributed through space and is likely to underestimate the real density of a species. I am therefore referring to my calculation as the relative density. For each grid i we have $p_i = 1 - e^{-\mu_i}$, where p_i is the proportion of grid i occupied (occurrences) and μ_i is their mean abundance within grid i, interpreted here as the relative density. This can be rewritten as $\mu_i = -\ln(1 - p_i)$. The occurrences, p_i, is calculated as $\frac{S_i}{R_i}$, where R_i is the number of record cards for grid i and S_i is the number of record cards that sighted European Starlings in grid i. The relative density of starlings in both SABAP 1 and SABAP 2 is shown in Figure 2.11.

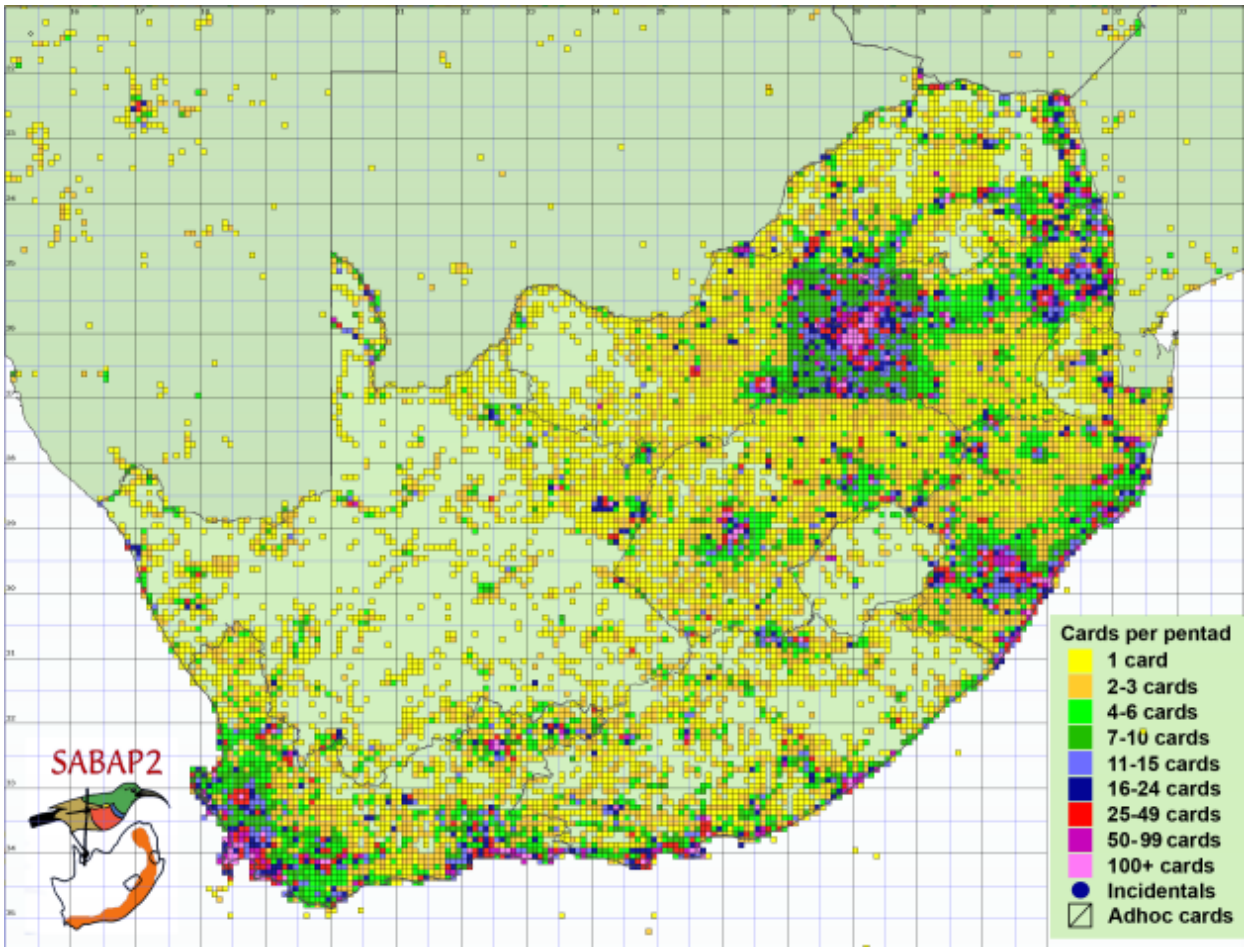


Figure 2.9: Number of checklists per grid cell of SABAP 2 since the beginning of data collection, 2007, up to November 2013 (ADU).

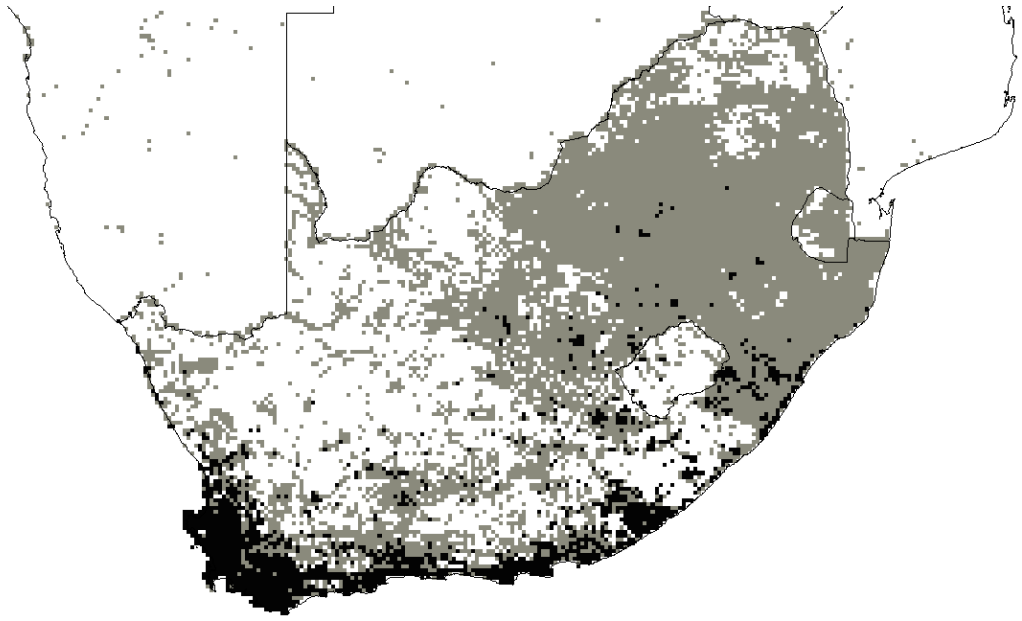
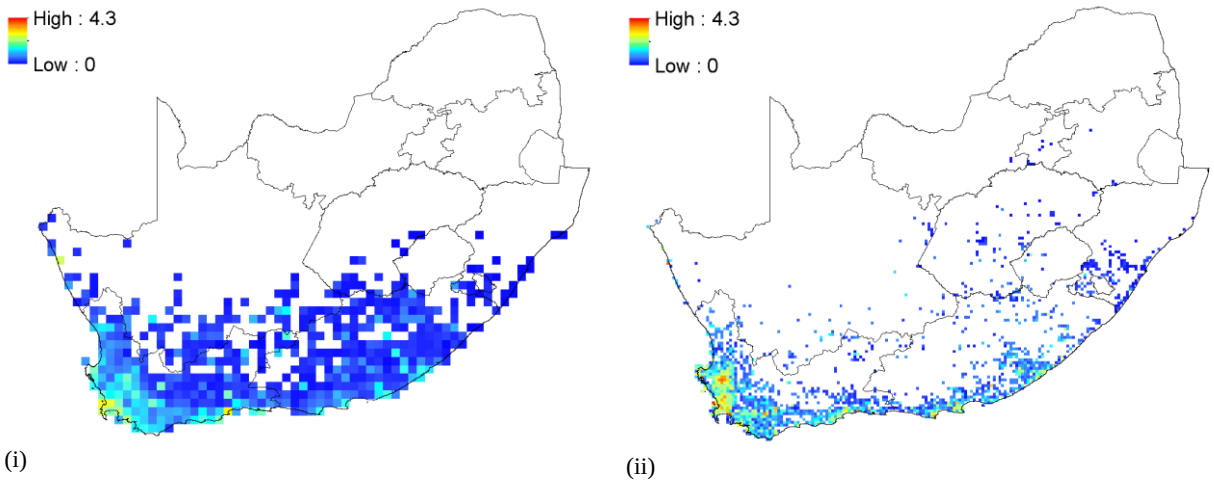


Figure 2.10: Reported presences of the European Starling (black) and absences (grey) as collected by SABAP 2 at 5 minute degree resolution since the beginning of data collection, 2007, up to July 2013. White areas indicate no data yet available.



(i) (ii)

Figure 2.11: (i) SABAP 1 and (ii) SABAP 2 relative densities for the European Starling.

CHAPTER 3

SPECIES DISTRIBUTION MODELS

3.1 Literature Review

The recent development of correlative species' distribution models (SDM, also known as climate envelope models, environmental/ecological niche models or habitat models) has made it possible to study the distribution of species across the earth. Predicting species' distributions is a focus point for many application studies in ecology, conservation and evolution science. SDMs are increasingly being used to address a variety of theoretical and ecological questions. These models allow us to make potential forecasts on how changes in climate and habitat could influence patterns of species' distributions. Furthermore, they contribute to our understanding of spatial patterns of biodiversity. This accumulation of knowledge could assist us in carrying out appropriate planning and management strategies.

SDMs are applied in marine, freshwater and terrestrial environments. These models have been used for various purposes, including but not restricted to (Guisan and Thuiller, 2005):

- quantifying a species' environmental niche,
- modelling species' assemblages,
- assessing the impact of climate and other environmental changes on species' distributions,
- assessing species' invasion,
- supporting conservation planning and reserve selection,
- supporting management plans for species' recovery and reintroduction sites,
- testing evolutionary, ecological and biogeographical hypotheses, and
- suggesting unsurveyed sites for rare species.

A SDM utilizes the relationship between species' occurrences and predictor variables (environmental and/or geographical) to define a species' potential distribution (Figure 3.1). Their expanding use and constant development means that these models are now able to accommodate a range of different data sets and different projection types, including climate modelling and the modelling of invasive species. Published studies have illustrated how well SDMs can characterise species' distributions in their natural ranges, especially when the data used is well formulated, an appropriate model is used with relevant predictors and the model is well evaluated. These models become more challenging to use as soon as we move away from their natural ranges into a different time or space, as is the case when studying invasive species.

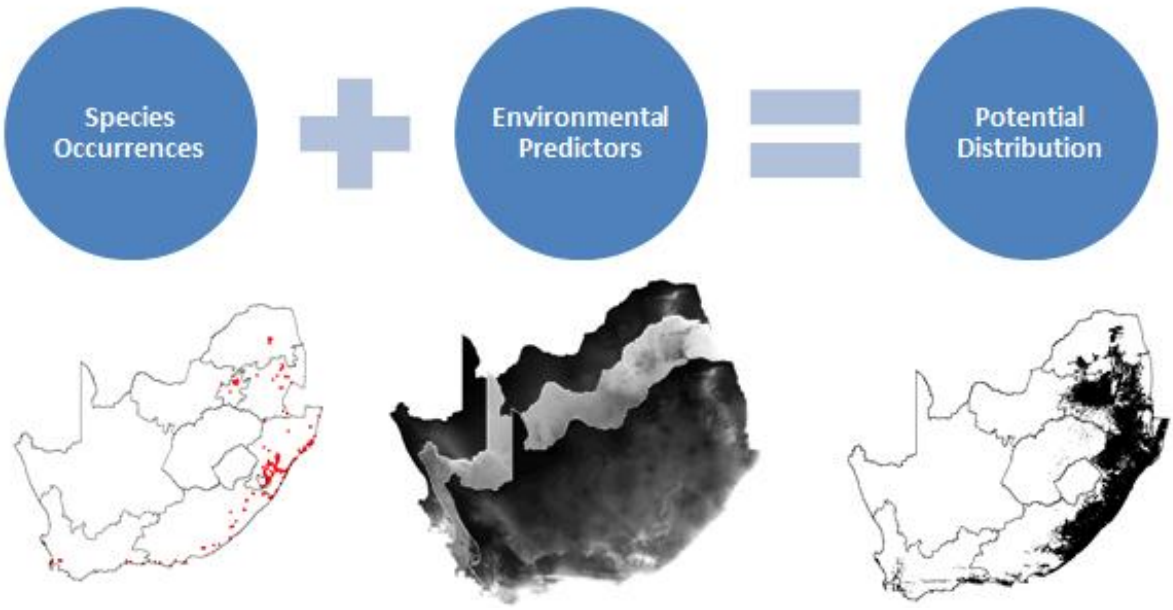


Figure 3.1: Schematic representation of a species' distribution modelling approach. The relationship between species' occurrences and environmental predictors is used to define the potential distribution of the species.

The species' occurrence data used for a SDM can be presence, presence-absence or abundance data obtained from natural history museum collections or from field sampling. Environmental predictors chosen can have different influences on the species; it can act as a limiting factor to their distribution, as a disturbance, or as a resource (Guisan and Thuiller, 2005). The data are usually sampled from a limited time and/or space, only giving us a one-part view on the distribution. This could indicate that a model fitting the observed distribution too tightly might be underestimating the true potential distribution of the species. An assumption is therefore made that the species are in equilibrium with their environment (Guisan and Thuiller, 2005). Very few studies critically consider how close their model system is to equilibrium or when a new equilibrium will arise (for example after the onset of an environmental change). An even greater problem arises when invasive species are studied under this assumption, as invading species are not in equilibrium with their new environment.

3.1.1 Species Distribution Modelling Methods

Modelling methods range from straightforward environmental matching models such as BIOCLIM and DOMAIN to increasingly complex models working with non-linear relationships such as Generalized Additive Models (GAM) and Maximum Entropy models (MaxEnt). Generalized Linear Models (GLM) were the initial regression based SDMs (Elith and Leathwick, 2009). Most SDM methods are regression-like, suggesting that additive combinations of predictors can model species' abundance. Multivariate Adaptive Regression Splines (MARS) use piecewise linear fits rather than smooth

functions, allowing for faster implementation than GAMs (Elith *et al*., 2006). Some of the initial SDMs only use presence data (BIOCLIM, DOMAIN, LIVES), but as SDMs developed most methods started to incorporate absence data as well, leading to an improvement in model accuracy. Machine learning and Bayesian methods are the most recent developments, allowing for sophisticated model fitting abilities. The complication is that these processes are more computationally intensive.

The recent advances in Geographic Information Systems (GIS) and spatial data technologies allowed for digital mapping, interpolation of climate variables and improved data collections and storage. Some methods are more focused on predictive modelling. These include the machine learning methods, Artificial Neural Networks (ANN), MARS, regression trees (eg Boosted Regression trees, BRT), genetic algorithms, support vector machines, and MaxEnt (Elith and Leathwick, 2009). These techniques are more complex and often viewed as 'black boxes', requiring greater insight into the ecological application and functioning of said techniques as well as limiting their use for those who lack the necessary expertise of computational systems.

Elith and Graham (2009) studied five methods with simulated data to compare and contrast the differences among the models and their performances. GLM, Random Forests (RF) and BRT were used for applying presence and absence data, while MaxEnt and Genetic algorithm for ruleset prediction (GARP) were used for presence only data. When studying the response functions (the relationship between species' occurrences and their environment), BRT and MaxEnt fitted the separate functions best while GARP performed the worse. For mapping the distributions, GARP over predicted, BRT performed best and MaxEnt second best, GLM and RF came close third. This is one of many studies illustrating the power and accuracy of machine learning algorithms such as MaxEnt and BRT (Elith and Graham, 2009; Elith and Leathwick, 2009; Elith *et al*., 2010; Pittman and Brown, 2011).

The choice of a modelling method should be determined by the available data, the intended research question and application thereof. No set of rules yet exist on which methods to choose when using different criteria and objectives. A few thoughts on this remain scattered throughout the literature without any definite guidelines available (Austin, 2007; Elith and Graham, 2009; Guisan and Zimmermann, 2000).

Model significance is influenced by the modelling method, the selection of predictor variables, the scale as well as the extent of extrapolation. Figure 3.2 illustrates the main steps necessary to obtain a well formulated model. If new knowledge is gained at the end of such a cycle, then the process can be iterated with appropriate adjustments to improve the model.

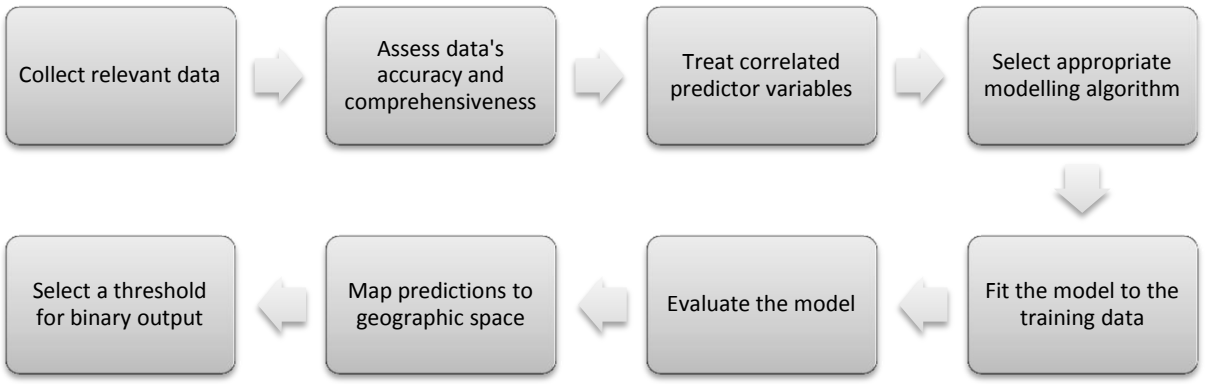


Figure 3.2: Schematic diagram of the key steps in implementing a species' distribution model (Elith and Leathwick, 2009)**.**

The modelling approach differs depending on the mobility of species. It tends to be easier to classify the environment of a sessile organism, whereas mobile species use resources that can be patchily distributed across a landscape making the modelling process more challenging (Elith and Leathwick, 2009). Access-related predictors which affect the species' ability to move around for example geographic barriers, waterways or road networks, need to be taken into account for highly mobile species. Additionally, observing mobile species and recording data are more challenging, which means that modelling techniques should try to allow for imperfect detection.

Choosing the correct spatial scale at which species' distribution modelling is undertaken is of utmost importance. Scale refers to both extent and grain (grids) and depends on the aim of the study, the available data and the ecological system; whether you are looking at global change (extent = global/continental) or regional conservation planning (extent = local). Grid size typically depicts the properties of the data or analysis, but also the characteristics of species' records and the intended application (Elith and Leathwick, 2009). When modelling climate impacts on a species' distribution for example, a macro scale study is more applicable than a micro scale study (Pearson *et al*., 2004). When focusing on micro scale, land cover plays a more significant role in a species' distribution than climate. Despite this segregation, it is important to find common ground where both land cover and climate can be incorporated into a model. Pearson *et al.* (2004) showed that better results were obtained when fitting their species' distributions after land cover data were incorporated in their distribution model.

Viewing and manipulating necessary data for the modelling procedure are best done in a GIS. During the procedure various aspects need to be considered, including multicolinearity, over fitting, over dispersion, etc. Limitations to reaching the full potential of SDMs come into play when ecological theory is not fully incorporated into the modelling process. Even though SDMs are ideal for studying various ecological theories, these theories should not be neglected throughout the process of building a SDM. Modellers

should use theories when selecting predictor variables, choosing ecologically realistic response curves, discussing the causes and costs of prediction errors, or when assessing the model's underlying assumptions (Guisan and Thuiller, 2005).

3.1.2 The Niche Debate

There is a controversial discussion among modellers on whether SDMs provide an estimate of the species' fundamental niche or realized niche. The fundamental niche argument is based on those models that only use climate variables to train the model. Only species-climate relationships are taken into account while leaving out other relationships such as competition and thereby only predicting the fundamental niche. Pearson *et al.* (2004) stated that the outcome of a SDM is an approximation of the fundamental niche of the species since the outcome is a suitable habitat only identifying areas where the species are likely to survive and reproduce. Guisan and Thuiller (2005) on the other hand stated that only mechanistic models that include physiological and behaviour parameters can model the fundamental niche.

The realized niche argument resides in the inherited constraints in the observed presence records that are used for model training. These records are samples from the realized niche where factors other than climate play a role in the species' distribution, therefore the outcome is also considered to be the realized niche. However, even though the species' data are constrained, it does not necessarily imply that the output will be the realized niche if no additional biotic interactions or barriers are incorporated into the modelling procedure. The majority of modellers lean towards the realized niche argument, even though some only assume their distribution is the realized distribution without giving sound ecological explanations for their reasoning (Guisan and Thuiller, 2005).

Invasive species often experience limited or no competition in their invaded range, making it more plausible to model their fundamental niche. Nonetheless, a SDM can not model their fundamental niche if only data from one of the invasive species' habitat regions are taken into consideration. It might be missing some important environmental features of suitable habitat in its native or other introduced regions where it has established, thereby limiting its range of suitable conditions and consequently limiting its fundamental niche. This argument could hold true for non-invading species as well. A species might have the potential to establish in areas beyond their current range, but if they have not had the opportunity to be introduced to such areas we lack the observational data of their existence in the greater environmental range. I can thereby conclude that we do not possess enough knowledge to even quantify a species' fundamental niche.

Hutchinson (1957) originally defined niches as follows: the fundamental niche is defined by the resources and positive interactions with other species that are required for a species to persist, while the realized niche incorporates the constraints that prevent a species from exploiting its resources. A simplified and straightforward view of a niche was defined by Chase and Leibold (2003) as "the environmental conditions that allow a species to satisfy its minimum requirements so that birth rate of a local population is equal to or greater than its death rate". It is possible that species' records, especially for invasive species, could have been recorded in an area where the individual was only 'passing through' and not necessarily have established there. Such records are misleading to niche modelling. Although this is worth taking note of, it is difficult to implement in field studies and distinguish between sexually reproductive individuals and non-sexually reproductive individuals only from single observations.

I will not add to the confusion of the niche debate and will from here onwards refer to the SDM output as a projection of the species' potential habitat/suitable habitat. Later on when combining the SDM output with my IBM where spatially explicit factors and population dynamics are incorporated, it can be referred to as a potential geographic distribution of the species.

3.1.3 Pseudo Absences

Due to the lack of absence data in most data collections (such as in museum records), some SDMs, including MaxEnt which I used, make use of 'pseudo-absence' records (also called background samples) to develop the model. Elith *et al.* (2006) completed an in depth study of 11 SDM methods and how well they predicted the species' distributions. Presence-only data with pseudo absences (a random sample of 10 000 sites from each region) were used for training the models and presence-absence data were used for evaluations. They found that the presence-only data were effective in modelling the species' distributions. It is possible that pseudo absences will coincide with presence records, especially when randomly choosing these pseudo absences from the study region. This is however widely accepted across different models and did not negatively impact the outcome.

Ideally data collections should strive to collect both presence and absence data, making the modelling more robust and accurate. Unfortunately this is not always possible and absence data could even be misleading due to species and their environment not being in equilibrium or the species might be difficult to detect. A few SDM methods (GARP, MaxEnt and Ecological Niche Factor Analysis (ENFA)) have been developed to deal with data sets lacking accurate absence data.

Studying the distribution of an invasive species with any SDM and defining their pseudo absences are challenging. The equilibrium assumption assumes that the current distribution of the species is a good indication of the conditions that the species can tolerate. However, this is not necessarily the case with an

introduced species. The species might not have had sufficient time since their introduction to become well established in their new habitat range and might still be in the process of exploring and expanding. Even though a specific area is therefore not occupied by the species it does not mean the area is unsuitable for them, which in turn puts a lot of constraint on the area available from which pseudo absences can be drawn. For this reason pseudo absences should be drawn from the invasive species' native regions as more information is often available regarding reliable absence records. On the other hand, in their new environment, invasive species tend to be released from their natural enemies from their native habitat. A niche shift may also occur where new environmental combinations are experienced to which they can adapt, possibly changing their geographic distribution in their introduced areas from their distribution in their native environment. The region defined for their pseudo absences will therefore depend on the stage of invasion. Initial introduced stages should require additional information from the species' environmental ranges in native regions, but as the invasion progresses and they've had ample time to establish and adapt, the introduced range could be able to provide enough information regarding suitable and unsuitable areas.

Defining the area from which the background points (pseudo absences) will be drawn is very important since this choice affects the outcome of the study. Here are some of the most important observations to keep in mind when choosing background points:

- Exclude areas where the species is barred by known geographic barriers (Elith *et al*., 2011; Phillips and Dudık, 2008).
- Exclude areas that have not been searched and only include background points from areas where it is known that sampling took place (Elith *et al*., 2011). This will allow a similar bias between presence and background points.
- Using randomly generated background points implies that the presences are also a random sample from all presence records, which is rarely the case (Elith *et al*., 2011).
- If a SDM is used for projecting outside the local geographic area, then using local background points will be limiting and prediction to other areas will result in substantial extrapolation (Elith *et al*., 2011).
- Background points that are too close to presences can give false projections. Likewise, if background points are too broad the projection can result in an over-prediction and a simplified model (Vanderwal *et al*., 2009).
- If projecting into a space where environmental conditions arise that were not involved in model training, unreliable projections is possible and the response curves might be incorrectly estimated. Thuiller *et al*. (2004) stated that projections of future distributions are only valid if the model is able to approach the complete response curve of the predictors.
- Presence points for another species that are part of the same survey as the focus species can be used (Phillips *et al*., 2006).

- Nakazato *et al.* (2010) created a buffer around the presence points for locally restricted and patchy populations.

Choosing which of the above methods for sampling background points is most appropriate depends on the knowledge and data available, the extent of modelling and the characteristics of the study area.

3.1.4 Predictor Variables

Great care should also be taken when choosing the predictor variables for a model as they are a primary decision maker of how the model output is formulated. It is important to choose variables that are relatively independent from each other, relevant to the dynamics of the study species and to the resolution of the study (Pearson *et al*., 2004). To build a meaningful model for a species' distribution you require knowledge regarding the species' biology, population dynamics, sensitivity to human disturbances, etc. Choosing more predictor variables would not increase the chances of a successful outcome (5 – 10 variables are considered ideal) and the balance of predictor variables should depend on the spatial scale being considered. Variables that have a direct impact on a species' distribution should be used above indirect variables, for example elevation, which has an indirect impact on the species' distribution.

It is essential to rely on *a priori* knowledge of which variables to include or exclude (Elith and Leathwick, 2009; Huntley *et al*., 2008). It is possible to start a SDM process with all the variables available and rely on the SDM's outcome to tell you the variables' contributions and accordingly eliminate variables, but this procedure can not replace one where a prior selection is built on existing knowledge and theory. On the other hand, if only *a priori* knowledge is used, a relationship is forced between the species' probability of occurrence and a climatic variable. When allowing the model to eliminate variables, additional relationships with previously unexpected variables which may be important for the species' distribution can be discovered, possibly leading to new knowledge regarding the species' habitat.

Studies modelling species' distributions often only consider climatic predictors. Depending on the species being studied, this can be a limitation to the model's efficacy and additional biotic processes needs greater consideration. Austin (2002) suggested that, when studying plants, processes such as dispersal, succession, grazing pressure, competition and fire should be incorporated in conjunction with climatic predictors. It becomes challenging when introducing variables such as competition to the model and distinguishing whether species' absence in a region could be due to climate or competition with another species.

Elith *et al.* (2008) used a step by step elimination strategy for simplifying their predictor set using BRT. After every step they dropped the least important predictor and then refitted the model again and continued doing so until a stopping criterion was reached. They ran these models with a 10-fold cross validation (CV) procedure and used the average CV error to decide how many variables could be removed before a reduction in predictive performance exceeded a predefined threshold.

3.1.4.1 Multicollinearity

High correlation between predictor variables will result in inaccurate regression coefficient estimates (of predictor variables) and therefore unreliable model outcomes. Most SDM methods do not discern between multicollinearity between variables, but it is necessary to confirm beforehand whether any high correlations exist. When this is the case, removing one of these variables would eliminate high correlations.

Collinearity may involve more than two predictors; it is therefore important to analyse pairwise correlation between all the predictors and Variance Inflation Factor (VIF) is the preferred method for studying correlation (Guisan and Zimmermann, 2000). The VIF is calculated for each continuous predictor as the regression of predictor i on the remaining predictors (R_i^2). This R_i^2 represents the proportion of variance in the i-th variable associated with the other variables in the model. Tolerance for the i-th variable is $1 - R^2$, the proportion of variance in the i-th variable not related to the other variables. VIF is the reciprocal of tolerance, $\frac{1}{1-R^2}$, the effects of R^2 on the variance of the estimated regression coefficient for the i-th variable (O'brien, 2007). The VIF estimates in other words how much multicollinearity has increased the variance of a slope estimate (Stine, 1995). The square root of the VIF tells us how much larger the standard error for the coefficient of that predictor is, compared with what it would be in a completely uncorrelated scenario. If all the VIFs are 1 it means that all the variables are orthogonal among themselves (Hugueny, 1989).

Deciding on a threshold value for VIF above which collinearity is high enough that a variable needs to be removed has been debated throughout literature (Craney and Surles, 2002; Elith *et al*., 2006; O'brien, 2007; Stine, 1995). No official threshold value exists, but commonly used threshold values are between five and ten. For this study I will use five as a threshold value for VIF. This relates to a R^2 value of 0.8. If VIF is higher than ten, Hugueny (1989) suggested that a more suitable method such as ridge regression or forward stepwise selection by the least-squares method be used instead.

3.1.5 MaxEnt

MaxEnt is a machine learning algorithm that estimates species' distributions by finding the distribution of maximum entropy, the distribution closest to uniform (Phillips *et al*., 2006). The species' occurrences are treated as samples from an unknown probability distribution across the study area. MaxEnt aims to estimate this unknown probability distribution. The known environmental conditions at these occurrences are one of the constraints set on the distribution. The sample average of the environmental variables (features) at the occurrences is taken. Consequently, the distribution of maximum entropy is such that for all features, *f*, mean(*f*) = sample average of *f*. However, the sample average is never exactly equal to the true mean, so a confidence interval is built around the sample average that will include the true mean. This is the second set of constraints (regularization). The final distribution of maximum entropy is thereby found such that the mean(*f*) falls within the confidence region of the sample average of *f*. The maximum entropy method applies these constraints to remove external influences and as a result the modelled species spreads out to fill only areas with suitable conditions that lie within the confidence region of the sample average.

Regularization is an approach to smooth the distribution through shrinking/penalising the coefficients to allow for a less complex and more general prediction (Elith *et al*., 2011). A small regularization parameter (less than 1) will over fit the model such that the predicted distribution primarily consists of those areas where the sample occurrences are found. The greater the regularization parameter, the larger the confidence interval and hence the smoother the response curves become, allowing for a more generalized distribution. Enforcing smoothness and a less complex model in this way has been shown to be misleading, especially when studying species' distributions in novel climates (Elith *et al*., 2010). It could lead to unrealistic predictions of equal probabilities in divergent environments.

When projecting into a novel climate it is probable that there could be variable values, not present during model training, for which predictions need be extrapolated. Such extrapolations should be treated with care. MaxEnt has a technique called 'clamping' for dealing with values that fall outside the training range. These values are treated as if they are the limit of the training range, remaining constant at those points outside the range. MaxEnt furthermore examines the extent of extrapolation by calculating multivariate environmental similarity surfaces (MESS) (Elith *et al*., 2010). It measures the similarity of the point to the distribution of reference points from the training range. Negative values are given to dissimilar points and maps are drawn for each point in the projection region. These are some features that distinguish MaxEnt from other SDM methods.

3.1.6 Boosted Regression Trees

As the name suggests, boosted regression trees (BRT) consists of two algorithms: boosting and regression trees. The former builds and combines models, while the latter forms part of the classification and regression tree group models. Boosting is a numerical optimisation technique used to overcome the inaccuracies inherent in a single tree model (Elith *et al.*, 2008, 2006). Improvement on a single model performance is accomplished by adding a tree at each step that reduces the deviance (error) of the previous model through applying the previously gained information, making small modifications to achieve a better model fit and combining the multiple models for prediction. Regression trees are good for selecting relevant variables and modelling interactions. Some studies have shown a preference towards BRT and demonstrated its ability to outperform other methods (Bahn and McGill, 2013; Elith *et al*., 2008, 2006). When implementing a BRT model you require presence data and absence/pseudo absence data. Both continuous and categorical predictor variables may be used. BRT can to some degree distinguish between correlated variables, making it a superior method for identifying important variables while at the same time taking into consideration multicollinearity.

3.1.7 Model Evaluations

Choosing the most appropriate method for evaluating a SDM depends on the subject and goals of the model. The model can range from a straight-forward descriptive model to a predictive model that requires greater complexity of model evaluation. Model verification assesses a model's ability to fit the training data, whereas model validation assesses a model's ability to predict events with independent test data (Araújo and Guisan, 2006; Elith and Leathwick, 2009). When simply trying to understand a model, model verification is appropriate and used mainly to verify the stability of the selected variables. Explaining patterns requires existing knowledge to make a comparison and statistical tests of model fit (Elith and Leathwick, 2009). When evaluating a model's predictive capability one needs to focus on the model's ability to predict independent events. This becomes complicated when modelling is done with the aim of predicting a species' distribution into a different scale, region or time period. There are a few statistical measures that modellers commonly use, including area under the receiver operating curve (AUC), Cohen's kappa statistic of similarity (k) and correlation coefficients (Elith *et al*., 2006; Pearson *et al*., 2004). The validity and sensitivity of these tests to data characteristics remain vague.

Despite doubts around its effectiveness, AUC remains one of the most widely used and unbiased measures of accuracy (Pearson *et al*., 2004). The AUC is obtained from the receiver operating curve (ROC), which depicts the relationship between the proportion of true positives (sensitivity) and false positives (1-specificity) with varying probability thresholds. It thereby measures the ability of predictions to discriminate between presences and absences (Elith and Graham, 2009). Good model performance is

characterized by large areas under the ROC curves, hence a curve that maximizes sensitivity for low values of 1-specificity (Figure 3.3). AUC ranges from 0 to 1, 1 being a model with perfect discrimination between presences and absences and 0 to 0.5 suggests the model is no better than a random model.

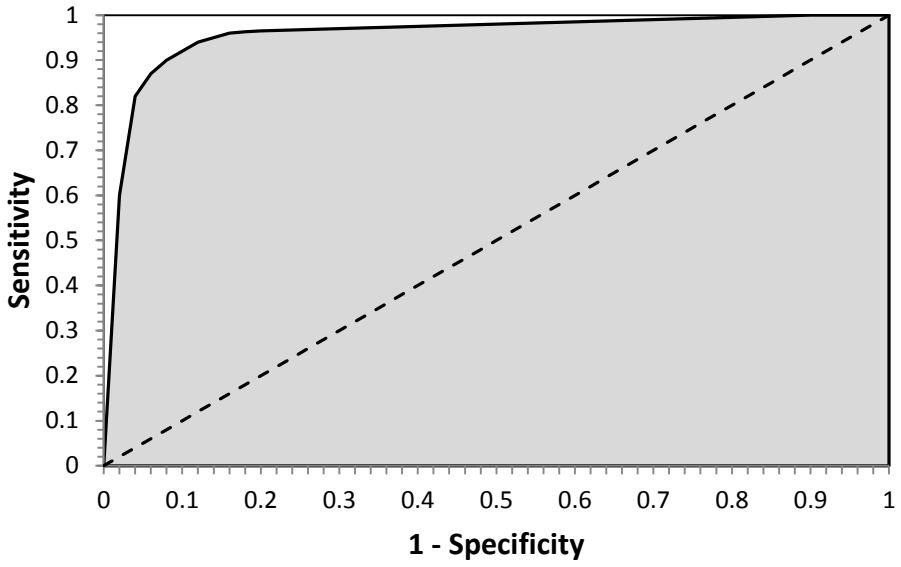


Figure 3.3: Area under the receiver operating curve (AUC, shaded region). The closer the curve gets to the upper left corner, maximizing sensitivity for low values of 1-specificity, the greater the AUC and the better the model's prediction of reality. The dashed line represents a random prediction (AUC = 0.5).

It is crucial to evaluate the model with independent data to ensure that over fitting does not occur. Over fitting would in turn lead to deceptive conclusions on the role of the predictor variables and their relationships with the study species. There are various ways of obtaining such a data set. The original data can be split into a separate training and test set using random split, spatial split or cross validation resampling methods. If independent data is available it can be used for testing. This data could be collected independently from the first data set, or it can be temporally or spatially independent data. Bahn and McGill (2013) concluded that a truly independent and spatially segregated data set is necessary to test if the model can be used to make predictions in new areas/environments.

3.1.8 Additional Modelling Challenges

Species distribution modelling is a field undergoing continuous development, but has already been used extensively. There remains however knowledge gaps and uncertainties regarding these models that need to be examined in greater depth for them to contribute with even greater satisfaction to conservation research. Araújo and Guisan (2006) identified areas of SDMs that require high-priority examination:

- Clarification of the niche concept (as previously discussed in Section 3.1.2).
- Improving sampling techniques for data collection. Due to SDMs sensitivity towards sample size and sample biases, sampling needs to be done in such a way that biases are limited and additional sampling is also required to have diverse and independent data sets.
- Improved parameterization strategies. There are many modelling techniques available to study SDM. It is not sufficient to only make comparisons between different techniques, but important to also look at within-model comparisons to understand the sensitivity of the models to certain assumptions and parameters. Variations in model outputs not only occur between different techniques, but also by varying parameters within the same technique. Such variations include: variable selection strategies, estimating absences in a specific way or the way spatial structures are considered. Knowing why different parameterizations provide different results is important.
- Improved model selection and predictor contribution.
- Improved model evaluation strategies.

Other challenging issues include the effect of spatial and temporal autocorrelation on models, the effect of geographical extent and resolution, the strategies for selecting pseudo-absences for model fitting (random versus selective) and rules for transforming modelled probabilities into presence-absences (maximizing the kappa statistic or using AUC index) (Araújo and Guisan, 2006). Accounting for biotic interactions is another challenge. It is a necessary predictor that needs to be considered, especially for extrapolation/forecasting. Additionally, the significant role that competition and other interactions play in shaping species' distributions might have great consequences for distributions, especially with climate change and novel environments.

From an ecology point of view there has been a lot of criticism towards SDMs for their lack of ecological theoretical grounding and sound ecological merit within the methods (Elith and Leathwick, 2009). Geographic and environmental processes play a fundamental role in species' distributions. Using both ecologically important variables and addressing geographic modelling is therefore essential when completing a SDM study. Suggestions are either to try and improve SDMs such that it better represents ecological processes, or combine SDMs with different models that do include such processes. The latter is what I did in Chapter 4 when I combined a SDM with an individual based model (IBM) to form a unique dynamic hybrid model.

3.2 Methodology

3.2.1 Objectives

My first objective was to find a potential distribution for the European Starling in South Africa which could then be implemented in my dynamic individual based model. SABAP 2 data were used as presence records. Due to the uncertainty regarding important predictor variables for the starling's distribution, two approaches were used for choosing predictor variables. The first approach involved *a priori* knowledge while the second approach involved a variable elimination procedure. Models were evaluated using AUC results along with other inspection techniques. MaxEnt was the main SDM method used for this purpose. Additionally, BRT and hierarchical partitioning assisted MaxEnt in the process of eliminating unnecessary predictor variables. The reasons why MaxEnt was preferred above all other methods available were because:

- I had presence-only data which is the type of occurrence data that MaxEnt uses,
- it can efficiently use both continuous and categorical environmental variables,
- it performed well when compared with other methods (Elith *et al*., 2006) and
- it is easy to use.

3.2.2 Occurrence Data

The SABAP 2 dataset was chosen for my presence records. The reasons why SABAP 2 was chosen above SABAP 1 were because of its finer grid size (5 minute versus 15 minute degree resolution) allowing me to study species-environment relationships at an appropriate scale for the European Starling, as well as it being the most recent data I had available regarding the European Starling's distribution in South Africa, giving me an updated indication of their range of environmental preferences in this introduced region.

The SABAP 2 presence records were processed in ArcGIS (ESRI, 2011) along with shapefiles of Southern Africa to ensure that there were no unexpected data records falling outside the continent's boundaries and that no duplicate records existed (Figure 3.4). The *x*-*y* coordinates of the SABAP 2 data points denote the upper left corner of 5 minute degree grids. The points were superimposed on the predictor variables and those points that fell outside a grid where variable data existed were removed.

(i) (ii)

Figure 3.4: Edited SABAP 2 (i) presence and (ii) absence records for the European Starling in Southern Africa.

3.2.3 Pseudo Absences

As discussed in the literature review (Section 3.1.3), many studies suggested the use of an invasive species' native absence/pseudo absence records when projecting to an invaded, non-native region. I decided not to use this method because (i) a possible niche shift took place and (ii) the starlings have had ample time (115 years) to establish in the new environment, giving me sufficient information regarding unsuitable areas where they have not settled. These areas, when carefully considering their accessibility, could be described as pseudo absences for the introduced region.

The European climate where the starlings naturally occur differ from the Southern African climate where they were able to establish successfully. This difference could result in inaccurate extrapolations when projecting to a climate where the model was not trained. I tested in MaxEnt the applicability of using European data for extrapolating to Southern Africa. Native European and invaded South African records were used for the presence sample while only European absences were used for the background sample. The projection is shown in Figure 3.5 (i). When comparing this projection to the actual observations of the starling in South Africa (Figure 3.5 (ii) SABAP 1 and (iii) SABAP 2), it was confirmed that a niche shift took place. Except for the small region along the Southern and eastern coastline, the rest of South Africa where the starling occur according to data from SABAP 1 and SABAP 2 are highly unsuitable according to the projection. This supported my decision to not use native records for modelling the starling's distribution in Southern Africa. The variables used for this modelling exercise, in order of variable contribution, were:

- mean winter temperature (58.8%),
- summer precipitation (23.4%),
- human footprint (8.7%),
- winter precipitation (6.6%),
- mean summer temperature (1.7%), and
- land cover (0.8%).

Winter temperature is clearly a defining factor for European distributions, which is not necessarily applicable to Southern African distributions.

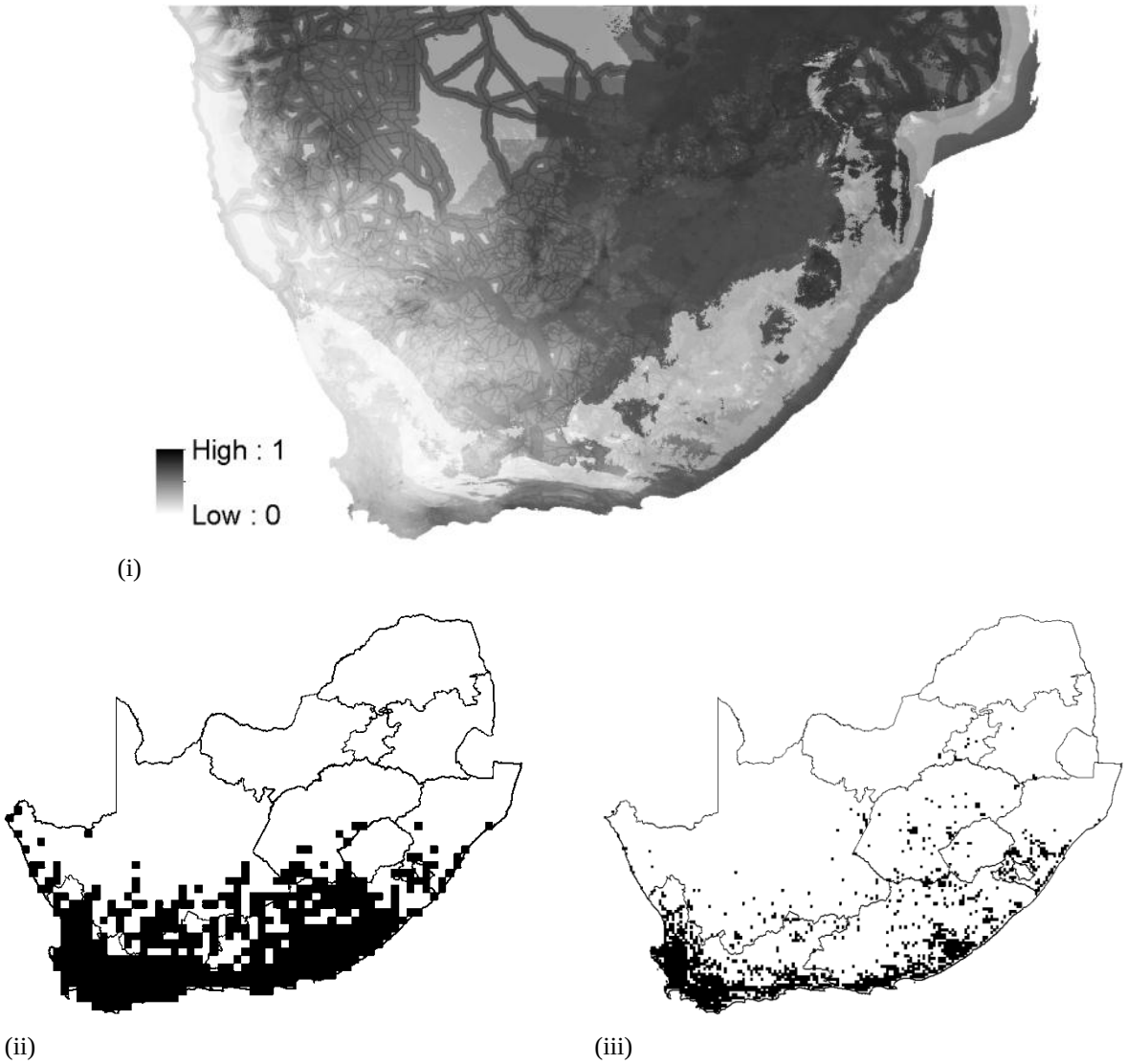
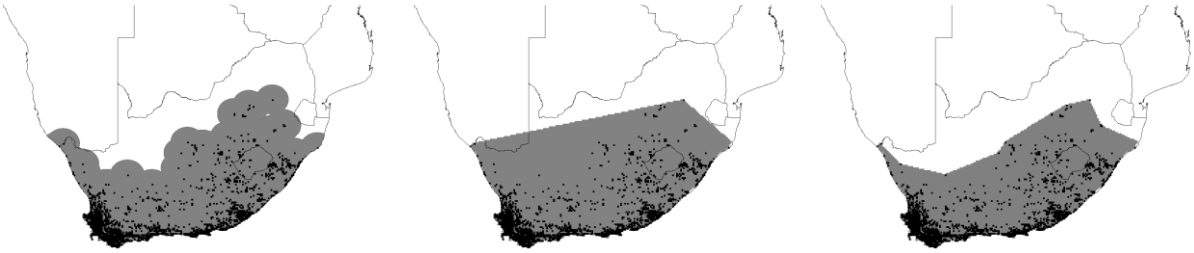


Figure 3.5: (i) A MaxEnt projection of the European Starling's potential distribution in Southern Africa at 1 km resolution (AUC = 0.662 (train), AUC = 0.630 (test)). Native European and invaded South African records were used for the presence sample, while only European absences were used for the background sample and 10% of the presence records were withheld for testing. (ii) SABAP 1 and (iii) SABAP 2 presence records shows the European Starling's observed distribution in Southern Africa.

In Section 3.1.3 I mentioned certain challenges when choosing appropriate pseudo absences (background points). My pseudo absences were therefore chosen to fill these knowledge gaps as much as possible. Since the starlings' range still continues to expand, I could not take any points outside their current range as pseudo absences. If they were absent there at the time of the study it did not infer that it was unsuitable habitat, as they could still establish there in the future. I constructed appropriate masked

regions that contained all the presence points (explained in more detail in the following paragraph and Figure 3.6). I thereby assumed that in the past 115 years since their introduction, the starlings have had enough time to spread throughout that specified region. Since the onset of SABAP 2's data collection process in 2007, 67% of South Africa has been covered when I collected my data from the ADU in 2012 (Figure 3.4). The remaining 33% of grids in South Africa were excluded from my background regions due to limited knowledge.

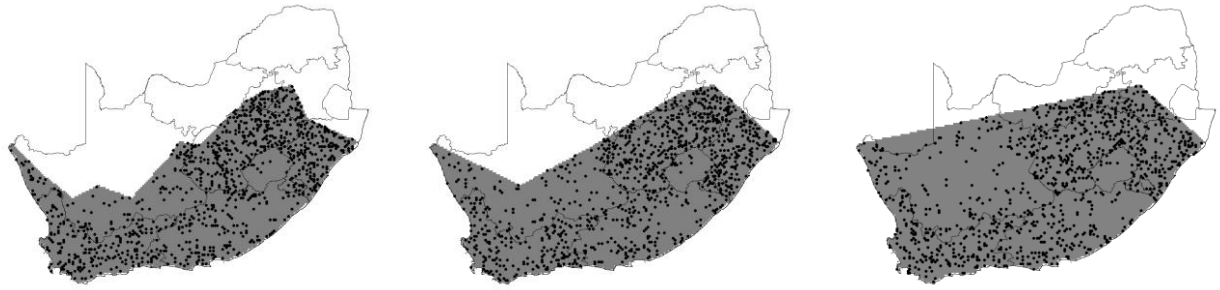
I created various masked regions in ArcGIS from which the pseudo absences were drawn. Buffers of varying kilometres (50 km to 200 km) were created around the presence records (Figure 3.6 (i)); a convex hull was created around all the presence points on the boundary (Figure 3.6 (ii)); and finally I aggregated points with 200 km to 500 km aggregation distances (Figure 3.6 (iii)). Aggregating points created polygon features around clusters of points within the specified aggregation distance (ESRI, 2011). For each point a polygon was created that included all the points within a radius of the specified aggregation distance (similar to the convex hull example). In the end all of these polygons were merged to form one polygon.



(i) (ii) (iii)

Figure 3.6: Examples of masked regions (shaded grey) created around the European Starling's presence records (SABAP 2, black points) from which the pseudo absences could be drawn. From left to right is (i) a buffer of 100 km created around the presence records, (ii) a minimum convex polygon created from the boundary presence records and (iii) aggregated points with an aggregation distance of 400 km.

Collections of pseudo absence points were created from sampling 1000 points from the SABAP 2 absence records, presented in Figure 3.4 (ii), which intersected each masked region. Examples are shown in Figure 3.7. I decided to exclude from my study those samples created by buffered regions around the presence records and the minimum convex polygon masked region. The reason for this was that they included areas beyond the starling's current distribution; this is visible in Figure 3.6 (i), (ii) and Figure 3.7 (iii). Those samples contradicted my statement that the background area should encompass a region in which the species had time to spread to. I was therefore only left with the aggregated points as masked regions.



(i) (ii) (iii)

Figure 3.7: Examples of pseudo absence collections where the shaded grey area is the masked region and the black points are the 1000 sampled absence records. The example masked regions are (i) aggregated points with a 300 km aggregation distance, (ii) aggregated points with a 500 km aggregation distance and (iii) minimum convex polygon.

3.2.4 Predictor Variables

A fine balance of predictor variables was needed, along with good reasoning why certain variables were chosen and others not. Both the spatial scale and species' requirements were considered when choosing predictor variables. No research had yet been published (by the time this study was completed) regarding appropriate environmental predictors for the starling's distribution in South Africa. In the first distribution model (SDM1, Section 3.2.4.1) I made my own postulates about important variables using *a priori* knowledge and in the second distribution model (SDM2, Section 3.2.4.2) I further investigated this research problem by conducting a variable elimination procedure.

3.2.4.1 Species Distribution Model 1

With the available knowledge of the European Starling and its habitat preferences, I designed SDM1 using predictor variables found to be relevant to the starling's distribution in South Africa. Various studies examined the distribution of the European Starling and other European birds in their native ranges and discussed relevant variables for their European distribution (Araújo *et al*., 2005; Araújo *et al*., 2009; Beale *et al*., 2008; Huntley *et al*., 2008). Many of these studies also considered the impact of climate change on the distributions and therefore chose variables that were more likely to change as climate changed. Huntley *et al.* (2008) used a range of variables that were related (directly/indirectly) to the primary climatic constraints (winter cold, growing season warmth and moisture availability) of species' distributions in Europe. These included the mean temperature of coldest month, the annual temperature sum above 5˚C and an estimate of the ratio of actual to potential evapotranspiration (this was estimated by taking daily values of precipitation, temperature and insulation as inputs). Table 3.1 presents other SDM studies of European birds and the respective predictor variables used.

Table 3.1: Examples of predictor variable combinations used in European avian SDM studies.

Beale *et al.* (2008)	Annual growing degree days	Mean temperature of coldest month	Soil water availability	Coefficient of variation in mean monthly temperature	
Beale *et al.* (2008)	Mean number of frosty days	Mean temperature of coldest month	Mean temperature of hottest month	Ratio of actual to potential annual evapotranspiration	
Beale *et al.* (2008)	Growing degree days	Mean temperature of coldest month	Mean monthly rainfall	Seasonal variation in rainfall	
Araújo *et al.* (2009)	Growing degree days	Mean temperature of coldest month	Annual precipitation	Ratio of actual to potential annual evapotranspiration	
Araújo *et al.* (2005)	Mean annual temperature	Mean temperature of coldest month	Mean temperature of hottest month	Mean annual summed precipitation	Mean sum of precipitation between July and September (growing season)

So far there has not been a great deal of studies examining bird distributions in Southern Africa using SDMs. From observations I knew that the European Starling is attracted to agricultural lands as a source of food, but also greatly attracted to urban settlements for both food sources and nesting spots. Their distribution was therefore highly likely to be influenced by human presences and disturbances, so I suggested the use of variables that exhibited these relevant characteristics, for example:

- land cover data that included cities, water bodies, croplands, grasslands, forests, deserts, etc.,
- anthropogenic biomes that covered dense settlements, villages, croplands, rangelands, forests, wild lands, etc.,
- human footprint data which gave the degree of human influence, or
- data illustrating the degree of land disturbance via crops/grazing.

I found various datasets on the above, but only one land cover and one human footprint dataset that satisfied my criteria (having an appropriate resolution of ≤ 1 km × 1 km grids, covered Southern Africa, consisted of well-defined categories and gave sufficient explanations regarding data collection). Global human footprint data were obtained from the Center for International Earth Science Information Network (CIESIN) (CIESIN, 2005) (Figure 3.8). This data was derived from the Human Influence Index (HII, produced by CIESIN in partnership with the Wildlife Conservation Society (CIESIN, 2005)) normalized by each biome. The HII covered population density, human land use, infrastructure and human access. It had a 30 arc seconds (30s) spatial resolution which is commonly referred to as '1 km' resolution (equivalent to about 0.93 × 0.93 = 0.86 km^2 at the equator and 0.83 × 0.83 = 0.69 km^2 in South Africa).

Land cover data was obtained from the Global Land Cover Facility (GLC2000, 2003), also at a 30s resolution, but with 0.89 km × 0.89 km grids (for South Africa) which encompassed 22 categories (Table 3.2 & Figure 3.9).

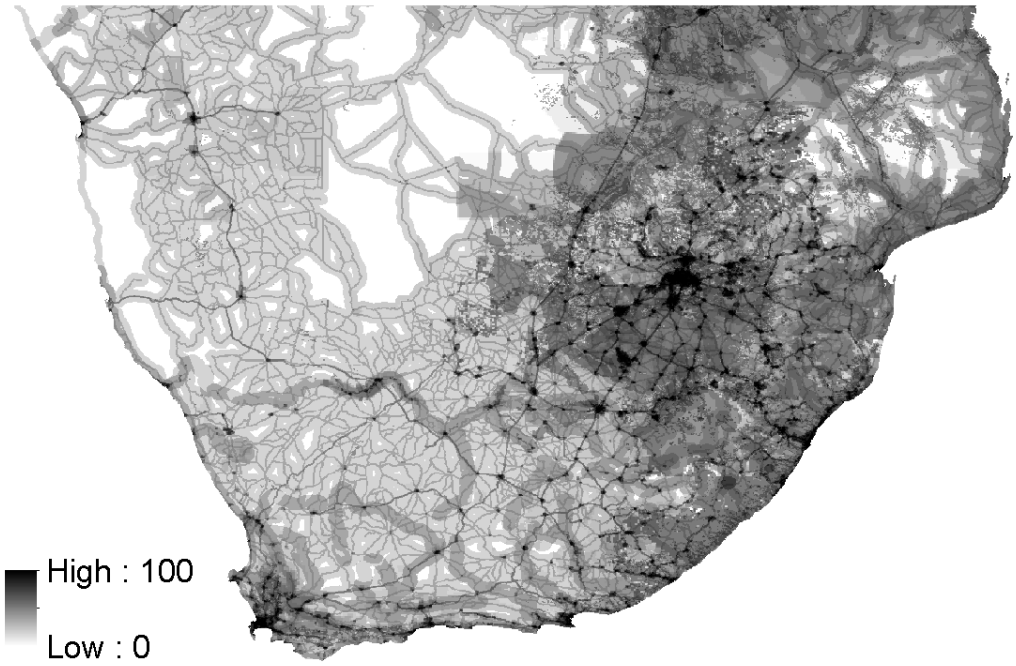


Figure 3.8: Human footprint data for Southern Africa represented at a 30s resolution.

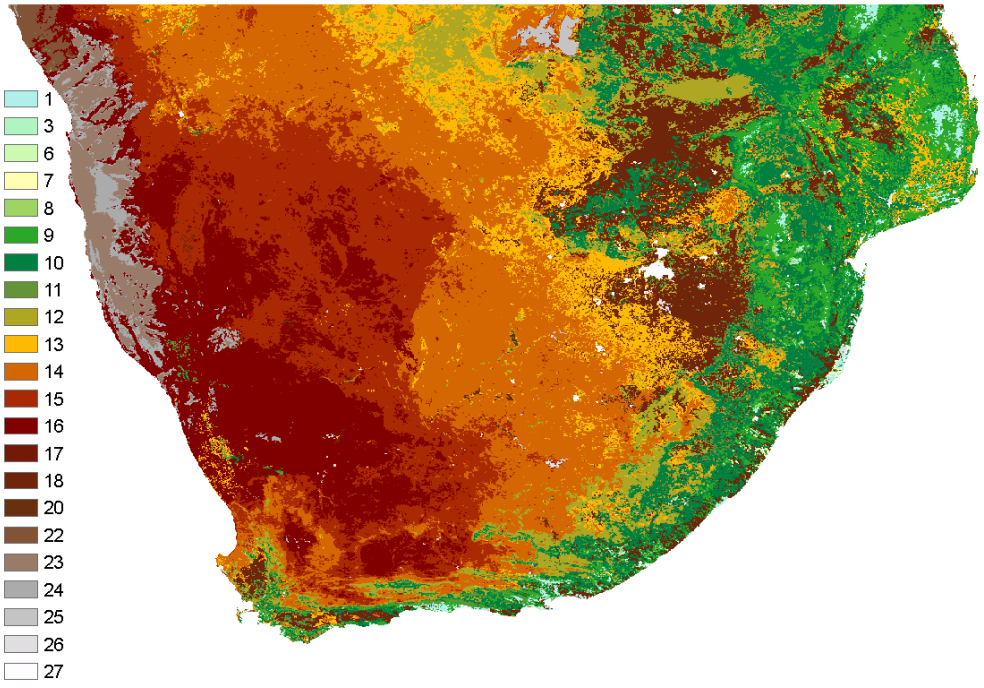
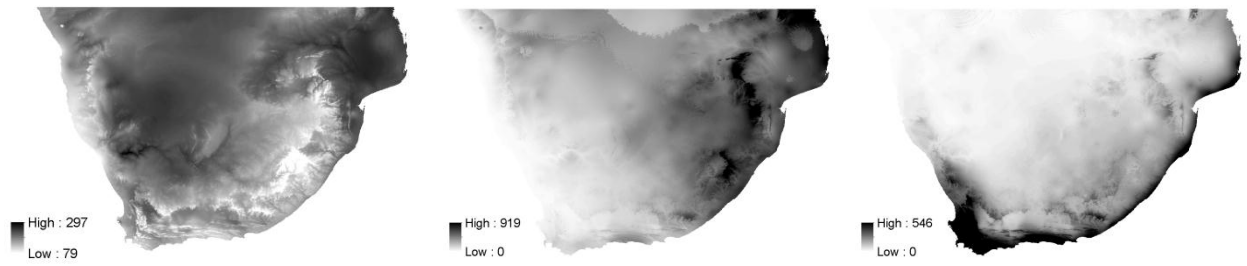


Figure 3.9: Land cover data for Southern Africa represented at a 30s resolution. The categories relating to the numbers in the legend are given in Table 3.2.

Table 3.2: Land cover categories as displayed in Figure 3.9.

1	Closed evergreen lowland forest	15	Open grassland
3	Submontane forest (900 -1500 m)	16	Sparse grassland
6	Mangrove	17	Swamp bushland and grassland
7	Mosaic forest / croplands	18	Croplands (>50%)
8	Mosaic forest / savanna	20	Irrigated croplands
9	Closed deciduous forest	22	Sandy desert and dunes
10	Deciduous woodland	23	Stony desert
11	Deciduous shrubland with sparse trees	24	Bare rock
12	Open deciduous shrubland	25	Salt hardpans
13	Closed grassland	26	Waterbodies
14	Open grassland with sparse shrubs	27	Cities

Other relevant factors that were considered for the starling's distribution in South Africa were moisture availability and temperature ranges (referring back to Table 3.1). The former could explain their limited distributions in the Karoo area observed in both SABAP 1 and SABAP 2 where rainfall was limited. Recent unpublished work (Berthouly-Salazar *et al*., 2013) done by geneticists on the European Starling in South Africa confirmed the importance of winter precipitation in the starling's distribution. This gave a genetic foundation for precipitation related selection in the starling's South African population. Researchers used a genetic approach called isolation by resistance where four alleles were identified of which three correlated with winter precipitation and the other one with both winter and summer precipitation. Contrary to the cold winters in Europe, cold temperature in South Africa might not be a limiting factor due to the warmer climate. Following the work of Beale *et al.* (2008) and Araújo *et al*. (2009), mean summer temperature was considered. The interpolated climatic variables were obtained from Worldclim (Hijmans *et al*., 2005) at a 30s resolution. The available bioclimatic variables are listed in Table 3.3. The choice of climatic variables included mean summer temperature (BIO10), summer precipitation (BIO18) and winter precipitation (BIO19) (Figure 3.10).



(i) (ii) (iii)

Figure 3.10: Bioclimatic variables chosen for SDM1 were (i) mean summer temperature (BIO10), (ii) summer precipitation (BIO18) and (iii) winter precipitation (BIO19). The temperature data is in degrees Celsius with a multiplication factor of 10. The precipitation is total precipitation for that quarter, given in millimetres. These variables are presented here for Southern Africa at a 1 km resolution.

Table 3.3: The list of bioclimatic variables obtained from Worldclim (Hijmans *et al*., 2005)**.**

BIO1	Annual mean temperature
BIO2	Mean diurnal range (mean of monthly (max temp - min temp))
BIO3	Isothermality (P2/P7) (* 100)
BIO4	Temperature seasonality (standard deviation *100)
BIO5	Max temperature of warmest month
BIO6	Min temperature of coldest month
BIO7	Temperature annual range (P5 P6)
BIO8	Mean temperature of wettest quarter
BIO9	Mean temperature of driest quarter
BIO10	Mean temperature of warmest quarter
BIO11	Mean temperature of coldest quarter
BIO12	Annual precipitation
BIO13	Precipitation of wettest month
BIO14	Precipitation of driest month
BIO15	Precipitation seasonality (coefficient of variation)
BIO16	Precipitation of wettest quarter
BIO17	Precipitation of driest quarter
BIO18	Precipitation of warmest quarter
BIO19	Precipitation of coldest quarter

The climate, land cover and human footprint data were all obtained from different sources and were edited in ArcGIS so that projections, grid sizes, grid alignment and spatial extent were consistent across all the variables. The data were all projected to the World Geodetic System from 1984 (GCS WGS 1984). This is a standard coordinate frame which is also widely used in Global Positioning Systems (GPS). When rescaling to bigger grids, the continuous data were resampled using cubic resampling where the weighted average was taken of the sixteen nearest points. When rescaling to a grid size only slightly bigger than the current grid size, bilinear sampling was used (this only took the four nearest points' weighted average). For the categorical land cover data I used majority resampling, which gave smoother results than using nearest neighbourhood resampling. Raster editing was done by keeping the processing environment fixed such that all rasters were snapped to a standard raster format and had the same extent as the standard raster. This ensured that all the variables' grid alignment and spatial extent were consistent.

Since SABAP 2 records were used as my sample points, all predictor variables needed to be in an appropriate resolution corresponding to that of SABAP 2 (5 minute resolution, about 8.333 km × 8.333 km grids in South Africa) in order to extract the predictor variables' values at all the sample points for both presence and pseudo absence records. The SDM projection was at the highest resolution possible to ensure the acquisition of detailed information regarding the starlings' distribution. Due to the discrepancy among the 30s resolutions (0.83 km versus 0.89 km) for the different variables, the chosen projection scale was 1 km × 1 km grids, which from now on will be referred to as 1 km

resolution. The climatic and human footprint variables were resampled accordingly. The extent of the projection output was intended only for studying the distribution along Southern Africa, so -20 degrees was the northernmost latitudinal cut off line. The final extent was -20˚S, 37˚E, -36˚S, 11˚E. All variables were clipped accordingly to this extent and converted to ASCII format for MaxEnt implementation. The final choice of variables depended on their degree of correlation. When multicollinearity, as measured by VIF, was above the chosen threshold level (VIF > 5) the relevant variable was removed to ensure no correlation existed between the chosen variables. VIF was calculated only among continuous variables.

3.2.4.2 Species Distribution Model 2

For the second modelling approach the research gap regarding the relationship between starlings and their environment in Southern Africa was investigated. A variable elimination procedure was designed to discover the most important variables that best explained the starlings' distribution, rather than only relying on inferences made on their habitat preferences as was done for SDM1.

Table 3.4 shows all the available variables that have been approved and edited so as to be used in SDM2. The land cover variable, previously used as a single variable, was now separated into its corresponding categories and used as individual continuous variables for the second modelling approach. This was done to study the contributions of each type of land cover to the starling's distribution in greater detail. The separation of the land cover variable was achieved using Geospatial Modelling Environment software (GME) (Hawthorne, 2012). The land cover raster was first transformed to 1 km polygons before separating each polygon in GME. The variables were then at a 1 km resolution with either 0 or 1 for each grid. When resampling to 5 minute resolution (for SABAP 2 data extraction), I took the percentage cover for each variable in a 8.333 km x 8.333 km grid, thereby ending up with continuous variables with entries ranging between 0 and 1.

The methods of maximum entropy (MaxEnt), boosted regression trees (BRT), hierarchical partitioning (HP) and variance inflation factor (VIF) were all used in a combined manner to eliminate irrelevant variables in a backward stepwise model selection process. Variables were eliminated one by one based on (i) biological relevance, (ii) multicollinearity and (iii) contribution to the species' distribution. VIF is highly sensitive to the combination of variables used, so when eliminating variables it was done first based on their biological irrelevance and then on their multicollinearity. The contribution of correlated variables was likely to be augmented, so the percentage contribution also changed when first removing variables with a high VIF. Therefore removal due to contribution was the final spot check.

Table 3.4: All 39 predictor variables implemented in SDM2, classified into climatic variables, land cover variables and other.

Climatic Variables	Land Cover Variables	Other
BIO1 Annual mean temperature	GLC1 Closed evergreen lowland forest	Human
BIO2 Mean diurnal range (mean of monthly (max temp – min temp))	GLC7 Mosaic forest / croplands	Footprint
BIO3 Isothermality (BIO2/BIO7) (* 100)	GLC8 Mosaic forest / savannah	
BIO4 Temperature seasonality (standard deviation *100)	GLC9 Closed deciduous forest	
BIO5 Max temperature of warmest month	GLC10 Deciduous woodland	
BIO6 Min temperature of coldest month	GLC11 Deciduous shrubland with sparse trees	
BIO7 Temperature annual range (BIO5-BIO6)	GLC12 Open deciduous shrubland	
BIO8 Mean temperature of wettest quarter	GLC13 Closed grassland	
BIO9 Mean temperature of driest quarter	GLC14 Open grassland with sparse shrubs	
BIO10 Mean temperature of warmest quarter	GLC15 Open grassland	
BIO11 Mean temperature of coldest quarter	GLC16 Sparse grassland	
BIO12 Annual precipitation	GLC18 Croplands (>50%)	
BIO13 Precipitation of wettest month	GLC20 Irrigated croplands	
BIO14 Precipitation of driest month	GLC22 Sandy desert and dunes	
BIO15 Precipitation seasonality (coefficient of variation)	GLC23 Stony desert	
BIO16 Precipitation of wettest quarter	GLC24 Bare rock	
BIO17 Precipitation of driest quarter	GLC25 Salt hardpans	
BIO18 Precipitation of warmest quarter	GLC26 Waterbodies	
BIO19 Precipitation of coldest quarter	GLC27 Cities	

MaxEnt and BRT were chosen due to them typically outperforming other methods in SDM studies (Elith *et al*., 2006). They were both implemented with a 10 fold cross-validation testing method. The pseudo absences were a sample of 1000 SABAP 2 absence records intersecting the masked region that was created by aggregating SABAP 2 presence points with a 500 km aggregation distance (Figure 3.11), the same background sample that performed best with SDM1 as results will show later). These two modelling methods were used for examining the percentage contribution (MaxEnt) and relative influence (BRT) of a variable before its removal or retention. HP was also used in this manner, but only once the elimination procedure has reached 12 variables, since this is the maximum number of variables implementable in the R package *hier.part* (Walsh and Mac Nally, 2013).

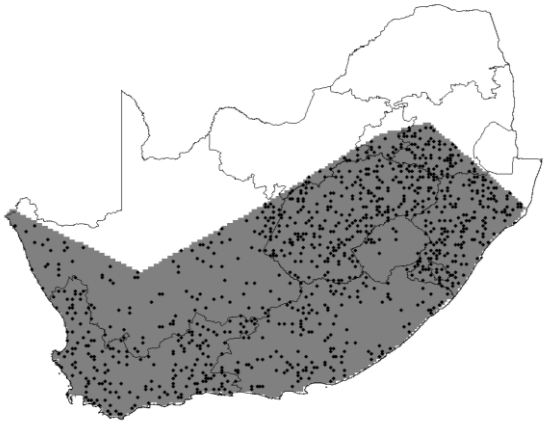


Figure 3.11: Masked region (shaded grey area) created by aggregating around SABAP 2 presence points with a 500 km aggregation distance. The points are a sample of 1000 points taken from SABAP 2 absence records that intersected the masked region. This sample was used as the pseudo absence points for the SDM2 modelling approach.

When a variable had the highest VIF, it was removed if and only if at least two of the three methods (MaxEnt, BRT and HP) showed a minimal contribution of this variable to the species-environment relation. If not the variable with the next highest VIF was inspected. Upon removal of a variable, the AUC (as provided by MaxEnt), was checked to confirm that no significant decrease (≤ 0.05) in model performance took place. This elimination procedure began with 39 predictor variables (Table 3.4) and stopped once:

- all VIF values were below 5,
- 10 or less predictor variables had been reached,
- the AUC had reached a maximum or decreased significantly with the removal of an additional variable, and
- BRT's simplification procedure lead to the result that the removal of the least influential variable would not improve the model's performance significantly.

For the very first elimination step I did a first round of inspection to see whether there were any variables which were highly irrelevant to my study species and/or study area that I could remove, knowing that removing it would have a negligible influence on the model results.

3.2.5 Modelling Setup

BRT was implemented in the R package *gbm* (Ridgeway, 2013) and the settings were similar to those implemented by Elith *et al.* (2006). MaxEnt version 3.3.3 was used and the settings were the same for both SDM1 and SDM2. These MaxEnt settings were the following:

- maximum iterations were increased to 5000 to ensure that the model has adequate time for convergence,
- the regularization parameter was set to 1, ensuring the model neither over fit nor under fit the distribution, but allowed for smooth response curves without risk of being too general, and
- the number of background points was set to 1000.

MaxEnt's projected output format was set to be logistic, the easiest format to conceptualise since it gave an estimate between 0 and 1 of the species' occurrence probability (Phillips, 2008). For BRT, the projection output format had to be normalized to obtain a range of occurrence probabilities between 0 and 1.

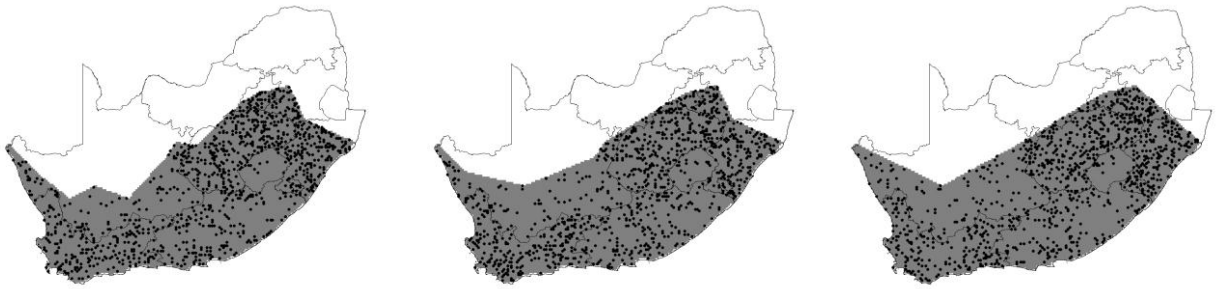
The data used for SDM1 testing were either a cross validation of 10-folds or an independent data set collected by Berthouly-Salazar *et al.* (2013) in May 2011. In this data set some points were clustered together in very close proximity (< 500 m). To avoid a conglomeration of similar predictor variable values, these clusters were edited such that only one central point remained.

I tested my distribution models' predictive ability since it was to be implemented in the IBM to assist in reconstructing the starling's range expansion process. For this reason I did not only use a 10-fold cross validation method for testing, but also an independent data set to test the model's true ability to correctly predict independent data points. I then compared my models' performances with each other based on the AUC statistic (referring back to Section 3.1.7).

3.3 Results and Discussion

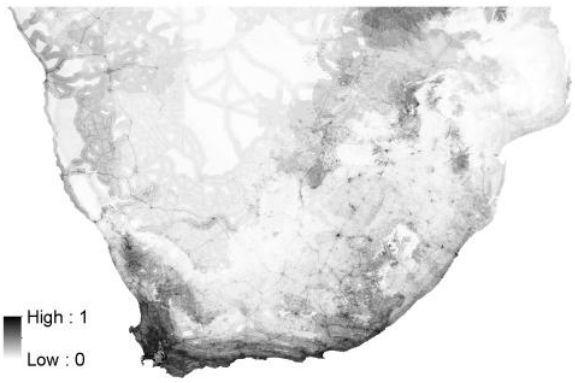
3.3.1 Species Distribution Model 1

All four chosen predictor variables (from Section 3.2.4.1, excluding the categorical land cover variable) had a VIF below 2 (BIO10 = 1.34, BIO18 = 1.66, BIO19 = 1.41, Human footprint = 1.23). Consequently all of the variables were used for further modelling since no significant multicollinearity existed among them. Samples of 1000 SABAP 2 absence records were taken from various aggregated point masked regions (with aggregation distances of 300 km, 400 km and 500 km around SABAP 2 presence records; Figure 3.12). The testing data varied between a 10-fold cross validation (CV) test and the independent data set from Berthouly-Salazar *et al.* (2013) (from here on referred to as the May 2011 data). Various model combinations were implemented in MaxEnt using different pseudo absence samples and testing methods. The models were compared to each other according to their AUC, extent of extrapolation and variable contributions. The results are shown in Figure 3.13 and Table 3.5.

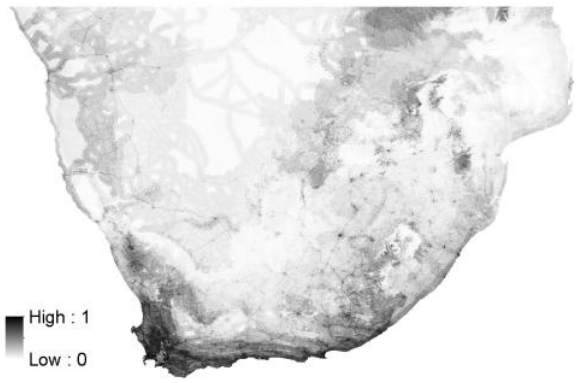


(i) (ii) (iii)

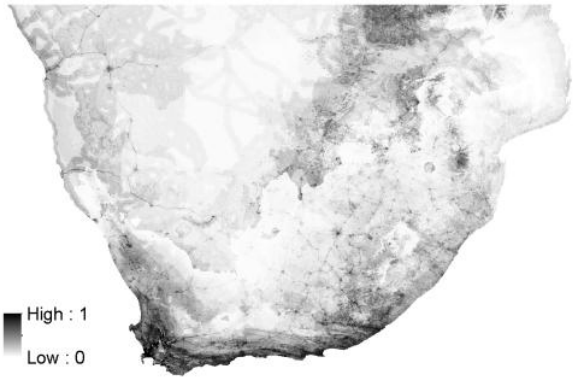
Figure 3.12: Pseudo absence samples used for SDM1. The masked regions are aggregated points around SABAP 2 presence records with aggregation distances of (i) 300 km, (ii) 400 km and (iii) 500 km. The points are a sample of 1000 SABAP 2 absence records.



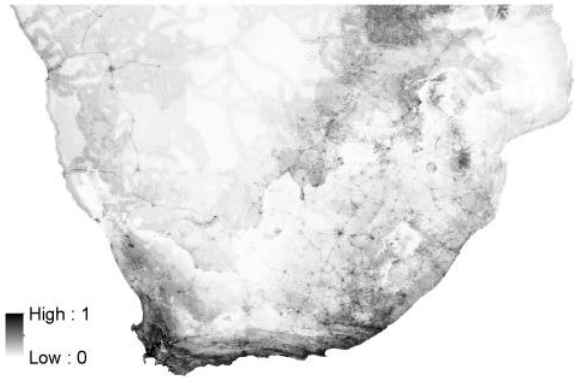
(i) **Model A**: Test - May 2011; Background – 300 km



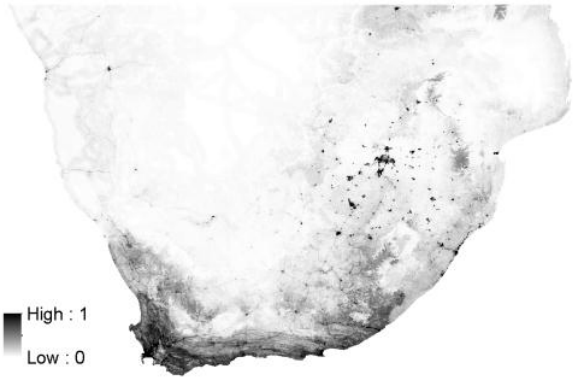
(ii) **Model B:** Test: 10-fold CV ; Background: 300 km



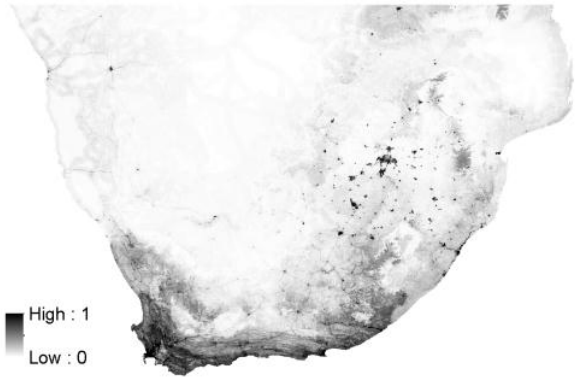
(iii) **Model C:** Test: May 2011; Background: 400 km



(iv) **Model D:** Test: 10-fold CV ; Background: 400 km



(v) **Model E:** Test: May 2011; Background: 500 km



(vi) **Model F:** Test: 10-fold CV ; Background: 500 km

Figure 3.13: MaxEnt projections of the European Starling's distribution in Southern Africa using various pseudo absence samples and testing methods. The different pseudo absence samples used were aggregated points with aggregation distances of 300 km (i and ii), 400 km (iii and iv) and 500 km (v and vi). (i), (iii) and (v) were tested with the May 2011 data while (ii), (iv) and (vi) were tested using a 10-fold CV method.

Table 3.5: AUC and variable importance results from MaxEnt for the European Starling's distribution in Southern Africa using various pseudo absence samples and testing methods. BIO19 = winter precipitation, BIO18 = summer precipitation, BIO10 = mean summer temperature, HFP = human footprint, GLC = global land cover.

	Test Data	Background Sample (aggregating points)	AUC	Variable Contributions (percentage contribution)
Model A	May 2011	300 km aggregation distance	0.836 (train) 0.893 (test)	BIO19 – 93.2 BIO18 – 2.7 HFP – 2.6 BIO10 – 0.9 GLC – 0.5
Model B	10-fold CV	300 km aggregation distance	0.826 (mean)	BIO19 – 89.2 GLC – 3.5 BIO18 – 3.2 HFP – 3.1 BIO10 – 1.1
Model C	May 2011	400 km aggregation distance	0.828 (train) 0.903 (test)	BIO19 – 76.3 HFP – 9.8 BIO18 – 9 GLC – 3 BIO10 – 2.1
Model D	10-fold CV	400 km aggregation distance	0.816 (mean)	BIO19 – 63.3 GLC – 14.1 HFP – 9.7 BIO18 – 9 BIO10 – 3.9
Model E	May 2011	500 km aggregation distance	0.838 (train) 0.908 (test)	BIO19 – 73.2 HFP – 10.9 BIO18 – 10.2 BIO10 – 4 GLC – 1.6
Model F	10-fold CV	500 km aggregation distance	0.825 (mean)	BIO19 – 63.1 HFP – 12.4 BIO18 – 9.4 GLC – 7.6 BIO10 – 7.5

The models performed similarly when measured by AUC, with Model E predicting a slightly higher AUC (0.838 for the training set and 0.908 for the testing set). The potential distribution maps (Figure 3.13) were also very similar, especially those for Model A - D, displaying high suitability for the European Starling in the Western Cape province and along the eastern coastlines. Intermediate probabilities were shown for the western coastline and inlands towards the Free State province, as well as for a large part of neighbouring Zimbabwe. Model E and F additionally displayed high suitability in the Gauteng province, which was evidently linked to the human footprint data as shown in Figure 3.8. This was also apparent in the contribution of the human footprint variable to these two distributions (10.9%

and 12.4%, respectively). Strangely enough, human footprint also contributed similar proportions to the distributions of Model C and D (9.8% and 9.7%, respectively), but the distribution maps themselves are analogous to that of Models A and B. Models A and B predicted very high contributions of winter precipitation (BIO19, 93.2% and 89.2%, respectively). All the models were in agreement that winter precipitation contributed the most to their respective distributions (BIO19 contribution > 60% for all models).

The background regions of 400 km and 500 km aggregation distances (Models C - F) were nearly equivalent (Figure 3.12 (ii) and (iii)), explaining their similar variable contributions. The background region of 300 km aggregated distance (Figure 3.12 (ii)) was more limited, which could explain the high contribution of only winter precipitation. This more restricted region also limited the amount of information received when training the model, leading to an increase in extrapolation and clamping (Figure 3.14 (i)). The interpretations of the predictions made in those areas were therefore questionable as these variables were outside the model's training range.

(i) (ii) (iii)

Figure 3.14: Clamping results for the three models tested with the May 2011 data set depicting where the prediction is most affected by variables outside their training range (lighter colours). The pseudo absence samples are aggregated points with aggregation distances of (i) 300 km, (ii) 400 km and (iii) 500 km.

Tests using the May 2011 data set performed better than the 10-fold CV method for all models as measured by AUC. The most likely reason for this was that all the 1395 SABAP 2 presence records were used for training the model whereas the CV method sampled 10% of the presence records for testing. Hence 10% less information was made available for training the model, leading to a model that was likely to perform worse. All three models A, C and E were able to predict distributions that performed very well when tested with the independent data set May 2011 (AUC $\geq$ 0.89).

Model E was chosen as the best potential distribution to represent the starlings' suitable habitat in the IBM model due to its slightly higher performance according to AUC and the limited amount of clamping present (Figure 3.14 (iii)). This model's various response curves are shown in Figure 3.15. Human footprint (Figure 3.15 (v)) was strongly positively related to the starling's presence probability. Summer precipitation (Figure 3.15 (ii)) tended towards a negative relationship with the starling's presence probability, while mean summer temperature and winter precipitation (Figure 3.15 (i) and (iii), respectively) exhibited a non-linear relationship. In terms of land cover types, a high likelihood of

presence areas corresponded predominantly to water bodies and deciduous forests (Figure 3.15 (iv) category 26 and 9-11). The land cover type least likely to be occupied by starlings was closed grasslands (Figure 3.15 (iv) category 13). All land cover types nonetheless predicted a presence probability greater than 50%, but as the land cover variable only contributed 1.6% to the starling's distribution, these values did not necessarily hold much significance.

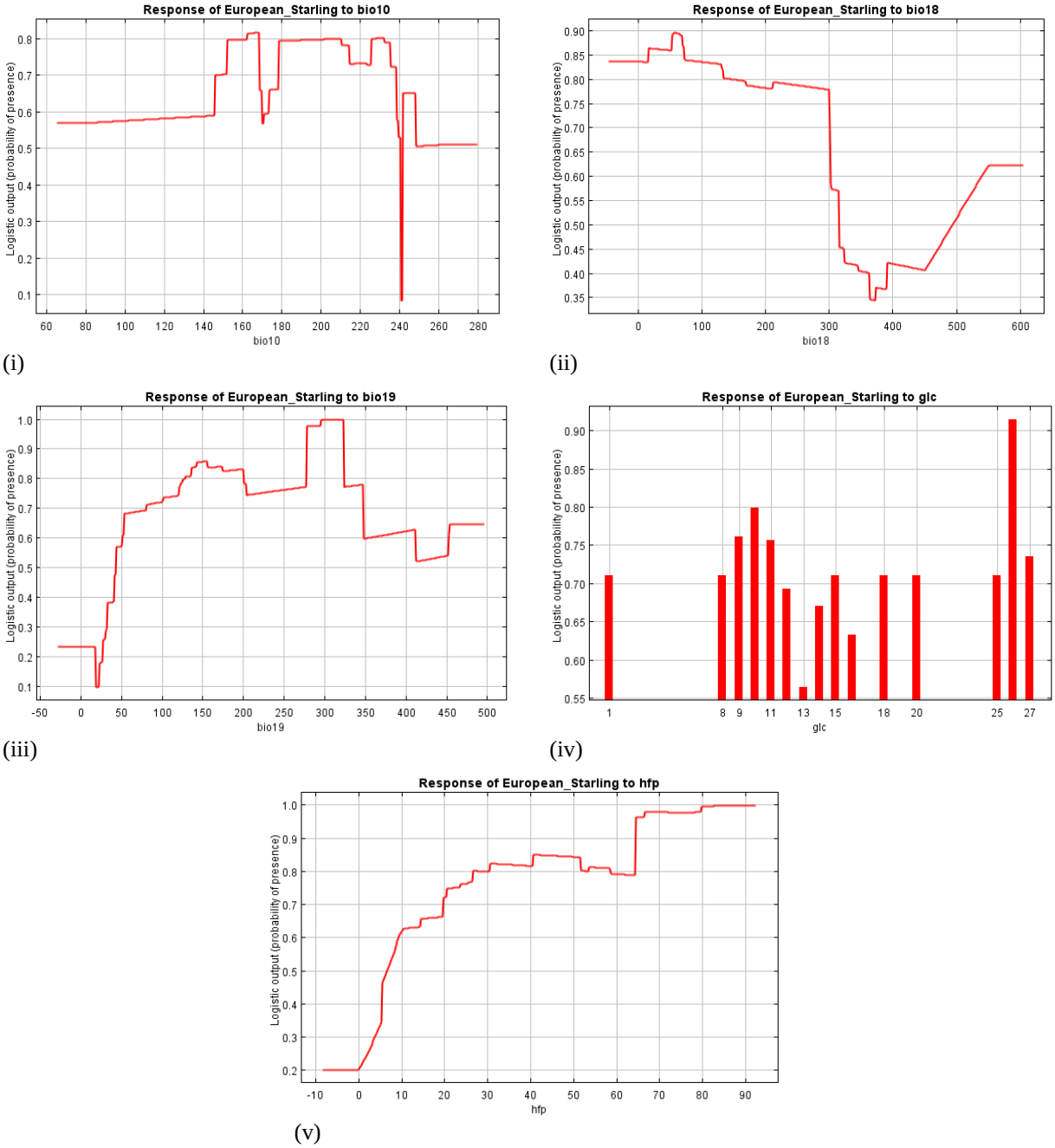


Figure 3.15: Response curves for the chosen SDM1, Model E, displaying the probability of starling presence for each variable's range of values. The variables are (i) summer temperature (BIO10, ˚C×10), (ii) summer precipitation (BIO18, total precipitation in mm), (iii) winter precipitation (BIO19, total precipitation in mm), (iv) land cover (GLC, categories as mentioned in Table 3.2) and (v) human footprint (HFP, %). The curves show how the logistic prediction changes as each variable is changed, keeping all other variables at their average sample value. Negative values are only due to MaxEnt performing clamping on those values at the end of the ranges as an extension to the available range of values.

3.3.2 Species Distribution Model 2

The variable elimination procedure began with the 39 predictor variables shown in Table 3.4. The very first step was to eliminate irrelevant variables by inspection. The following land cover variables were removed due to limited/no influences in my study region, specifically in South Africa where the SABAP 2 presence records were allocated (Figure 3.16): mosaic forest/croplands (GLC7), irrigated croplands (GLC20), sandy desert and dunes (GLC22), stony desert (GLC23), bare rock (GLC24) and salt hardpans (GLC25).

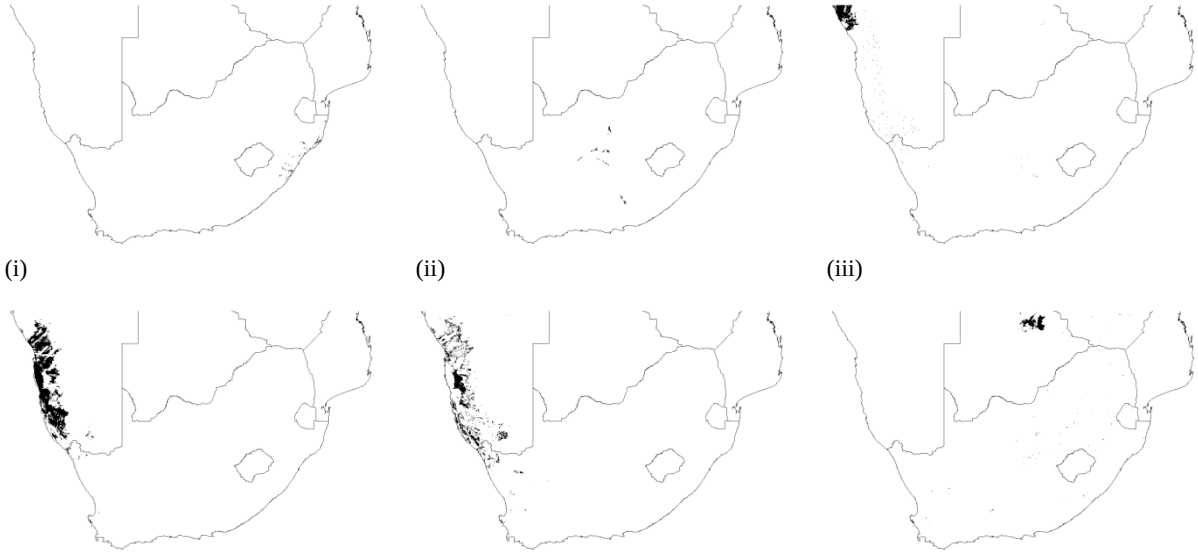
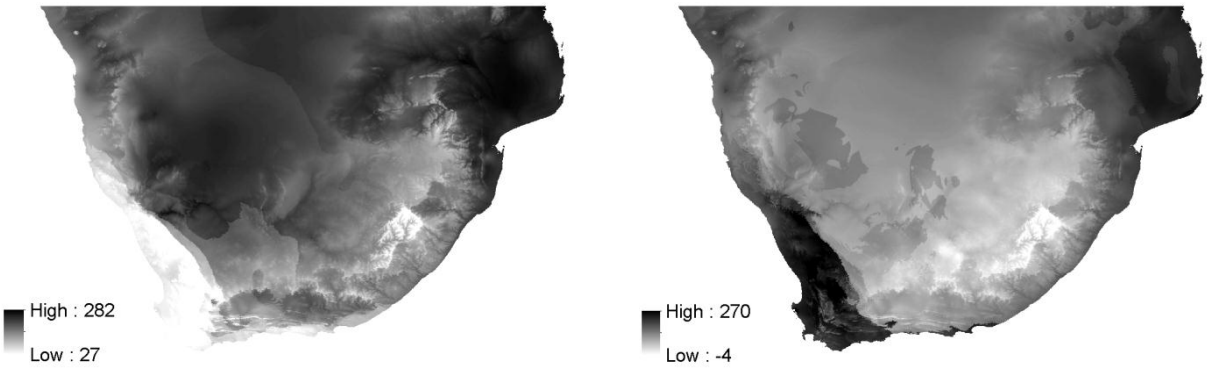


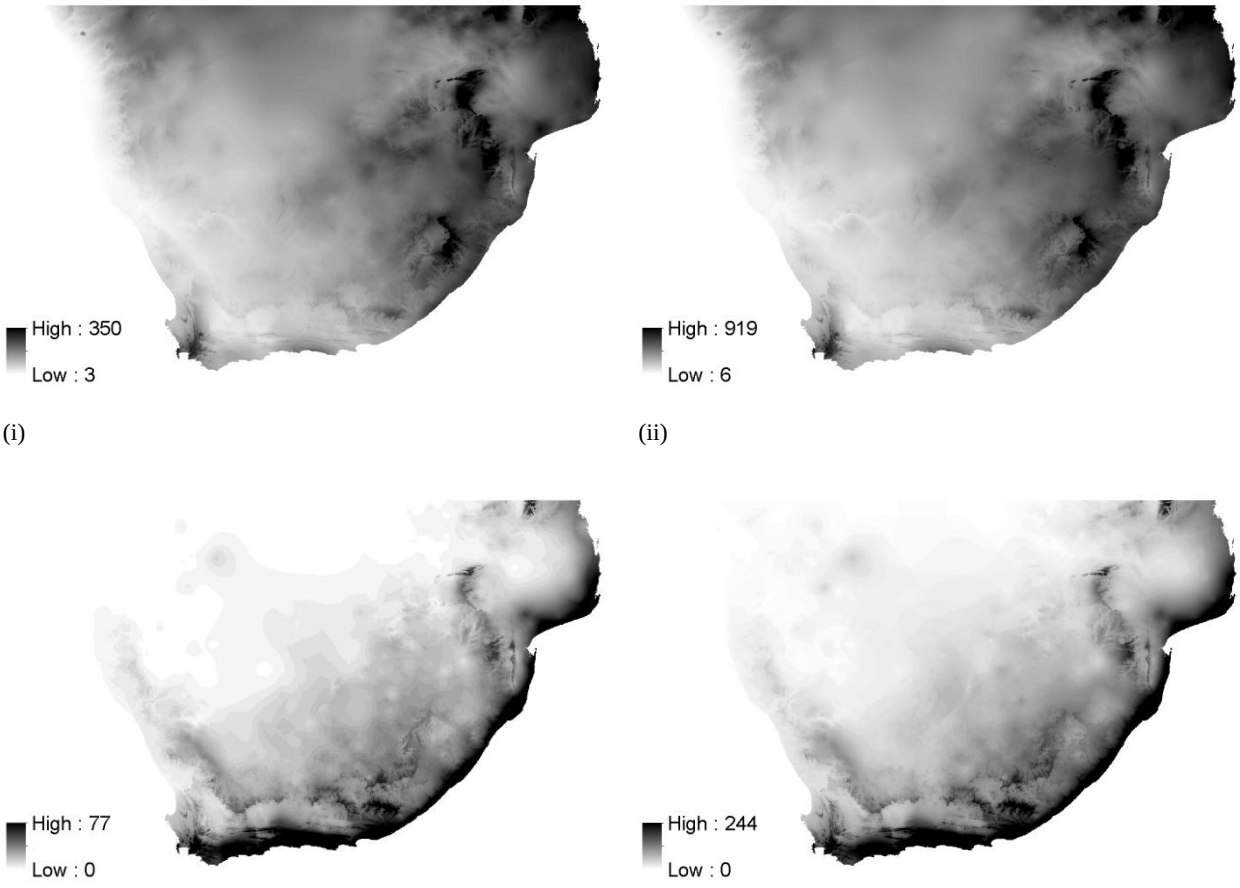
Figure 3.16: Land cover variables with limited coverage in South Africa that were removed during the first round of inspection of SDM2. These were (i) mosaic forest/croplands, (ii) irrigated croplands, (iii) sandy desert and dunes, (iv) stony desert, (v) bare rock and (vi) salt hardpans.

Climate variables that were eliminated by inspection were: mean temperature of wettest quarter (BIO8), mean temperature of driest quarter (BIO9), precipitation of wettest month (BIO13) and precipitation of driest month (BIO14). BIO8 and BIO9 were eliminated because of interpolated irregularities in the original data. These irregularities are evident in Figure 3.17 where uneven contours were observable between different regions. Such variables were unreliable and were left out as they could negatively impact the model results. BIO13 and BIO14 also portrayed irregularities, but to a lesser degree (Figure 3.18 (i) and (iii)). They were however nearly identical to BIO16 (precipitation of wettest quarter) and BIO17 (precipitation of driest quarter), respectively, which were observed to be smoother, so BIO16 and BIO17 were used instead.



(i) (ii)

Figure 3.17: Interpolated irregularities in (i) mean temperature of wettest quarter (BIO8) and (ii) mean temperature of driest quarter (BIO9), shown here at their original 30 arc seconds resolution. These variables were removed during the first round of inspection of SDM2.



(i) (ii)

(iii) (iv)

Figure 3.18: Comparing similar climatic variables and their degrees of irregularities. (i) Precipitation of wettest month (BIO13) and its similar counterpart (ii) precipitation of wettest quarter (BIO16). (iii) Precipitation of driest month (BIO14) and its similar but smoother counterpart (iv) precipitation of driest quarter (BIO17). All images are displayed at their original resolution of 30 arc seconds. BIO13 and BIO14 were removed during the first round of inspection of SDM2.

After removing the six irrelevant land cover variables and the four unreliable climate variables, 29 variables remained. Variables were then eliminated according to the steps set out in Section 3.2.4.2. This elimination procedure is illustrated in Appendix A and includes all the results of multicollinearity (VIF), AUC statistics and variable contributions from MaxEnt, BRT and hierarchical partitioning for each set of variables. When I reached 15 variables, all the variables' VIF were below two so I discontinued calculating the VIF. Throughout the process the AUC lingered around ±0.87 and then increased to 0.919 when nine variables remained, after which it decreased to 0.896 when only seven variables were left. After nine variables, BRT's simplification procedure led to the conclusion that the removal of the least influential variable did not improve the model's performance. The end criterions were all met and the distribution model with nine variables was chosen as my second potential distribution for the European Starling in Southern Africa (Figure 3.19). The nine variables were, in order of percentage contribution in MaxEnt, human footprint (44.9%), annual mean temperature (30.4%), deciduous woodland (4.8%), open grassland with sparse shrubs (4.7%), croplands (4.4%), winter precipitation (3.9%), closed deciduous forest (3.3%), summer precipitation (2.2%) and sparse grassland (1.3%).

A high suitability for the European Starling, similar to SDM1, was evident in the Western Cape province and along the eastern coastlines with an additional extension of high suitability areas along neighbouring Namibia's coastline and inlands towards the provinces of the Free State, KwaZulu-Natal, Gauteng and Mpumalanga. This model performed very well with an AUC of 0.919, possibly explained by the large proportion of very high suitability areas and the greater variety of variables allowing more modelling flexibility. Winter precipitation played a significantly less important role in this model than in the first distribution model (3.9% versus 73.2% contribution) while human footprint increased in importance (44.9% versus 10.9%). All four land cover variables contributed less than 5% to the potential distribution, but their influence could be seen nonetheless in the patchiness of SDM2 as oppose to the smoother distribution of SDM1.

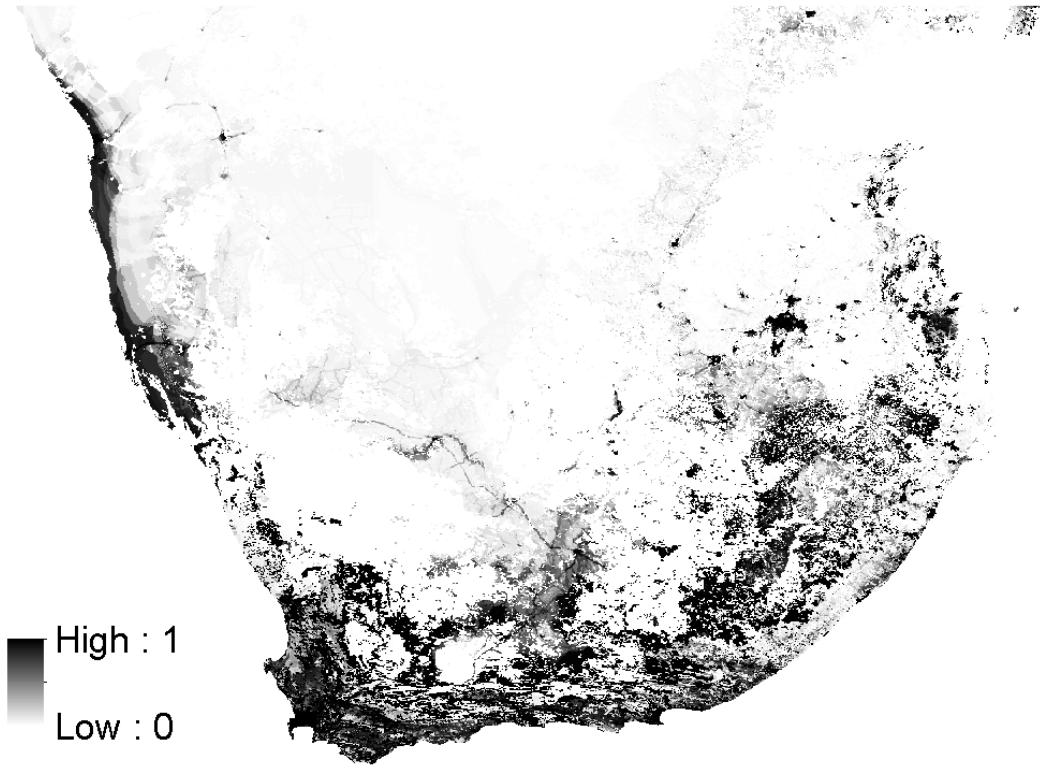


Figure 3.19: SDM2's potential distribution for the European Starling in Southern Africa, shown at a 1 km resolution. This is the average distribution for the 10-fold CV method.

3.3.3 Concluding Remarks

Studying species distribution models still pose a great variety of challenges as I have discussed in Section 3.1.8. Relating to my study, I have looked in great detail at choosing appropriate backgrounds samples as well as predictor variables. Interpolated data such as the climate variables applied here should be used with care as I have done when I removed those variables containing irregularities. Further attention could be given to more detailed parameterisation strategies and model evaluations. It is necessary to study the reliability of MaxEnt, as well as other methods, when slight changes are made to parameters such as the sample size of pseudo absences and the sample area. For my purposes I did not evaluate these models in greater detail as this is only a stepping stone to my modelling core which is the dynamic model. I have chosen two realistic distribution models that perform well when measured with AUC with which I will continue to the modelling core.

ROC has a tendency to favour conservative models. I saw this expressed predominately in SDM2, but also in SDM1, where the distribution prediction was limited to those areas where SABAP 2 presences occurred which resulted in a high AUC of 0.919 (Figure 3.20). For this reason using AUC as an evaluation technique is not necessarily the best method when studying invasive species. It is better to over predict than under predict the potential distribution of an invasive species, especially for

management and conservation purposes. If an area that could become invaded is under predicted due to a conservative model, appropriate management plans would not reach this area. This could lead to an unforeseen invasion that poses great risk which could have been avoided. When using MaxEnt, this problem can be overcome by setting the regularization parameter > 1. However, this fell outside the scope of the study. As my second objective stated, I wanted to create a dynamic model that could successfully reconstruct the range expansion of the starling in Southern Africa, and for this a conservative distribution map is sufficient as it portrays the suitable habitat closely related to their currently observed distribution.

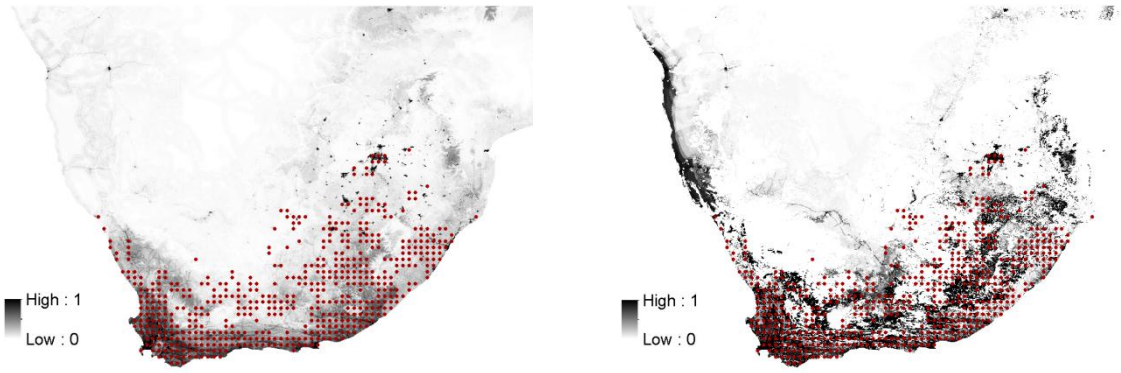


Figure 3.20: Potential distributions of the European Starling in Southern Africa as predicted by (i) SDM1 and (ii) SDM2. Shown in red are the SABAP 2 presence records used when building both these models in MaxEnt.

A SDM forces a relationship between the study species and the environmental variables. Their distribution is limited by the variables used. I know that there are other factors essential to a species' distribution that we cannot study with a SDM. In the words of Beale *et al*. (2008), "the degree to which species really are constrained by climate remains unresolved". It is a complicated relationship to understand, even with the knowledge and modelling techniques available. One way of improving on the restrictions posed by a SDM is by approaching other techniques such as mechanistic models to assist in explaining a species' distribution. I did this by using an individual based modelling approach where I incorporated a greater variety of processes.

CHAPTER 4

INDIVIDUAL BASED MODELS

4.1 Literature Review

It is challenging to numerically simulate complex biological processes such as the spread of invasive species. An individual based model (IBM) is one tool available to simplify such systems and study them computationally. IBMs are parallel to the more general agent based models or multi-agent systems where an agent either refers to an individual or a group of individuals. According to DeAngelis and Mooij (2005) no absolute definition exists for an IBM, but in broad terms an IBM simulates the actions and interactions of individuals within a population across space and time. It is a bottom-up approach for studying individuals in order to better understand properties of the system that emerges from these individuals and the interactions among them (Grimm, 1999). Each cell is comprised of a number of individuals from the study species and at each time step certain rules or state transition functions determine each individual's fate depending on its position in the previous time step as well as the state of the surrounding environment. An IBM incorporates differences among individuals such as differences in experience and learning, genetic variability, phenotypic variability, behaviour, life cycles, movement and local interactions (DeAngelis and Mooij, 2005).

The use of IBMs in ecology have rapidly expanded in the past two decades (Grimm, 1999). There is however no standard protocol for describing IBMs which often makes these models difficult to understand and reproduce, especially without any computational background. It is also more complicated to analyse, evaluate and communicate than analytical models. At present only a limited number of examples exist regarding the study of invasive species with IBMs (Higgins *et al*., 1996a; Jongejans *et al*., 2008; Nehrbass and Winkler, 2007; Travis *et al*., 2007).

4.1.1 Model Evaluations

Evaluating IBMs are often challenging due to the nature of the modelling approach. In this study I focused on evaluating them based on their goodness of fit using presence-absence as well as density data obtained from my IBM. Elith and Graham (2009) highlighted the importance of using more than one evaluation method for assessing model performance because of the different aspects of performance that each technique quantifies. There are a few statistical measures that modellers commonly use for evaluating presence-absence data. These are (Bahn and McGill, 2012; Elith *et al*., 2006; Richard Pearson *et al*., 2004):

- area under the receiver operating curve (AUC, already explained in Section 3.1.7),
- Cohen's kappa statistic of similarity (and other confusion matrix tests such as the true skill statistic), and
- correlation coefficients.

The validity and sensitivity of these tests to data characteristics remain vague. As with SDMs, it is crucial to evaluate the model with independent data to avoid misleading conclusions.

A confusion matrix, also known as a contingency table, can be obtained from comparing the presences and absences of both the model outcome and the observational data set. It is an evaluation method that allows one to visualize the performance of an algorithm. The outline of a confusion matrix is shown in Table 4.1. Well known measures such as overall accuracy, sensitivity, specificity, true skill statistic (TSS) and Cohen's kappa can then be calculated to assess the model's performance (Table 4.2) *(Liu et al*., 2009). Sensitivity is the proportion of correctly predicting observed presences (quantifying omission errors) while specificity is the proportion of correctly predicting observed absences (quantifying commission errors). On their own they are restricted to evaluating only one aspect of a model's predictive ability, but when combined they can be more efficiently used as an evaluation technique similar to what is achieved by AUC and TSS. Overall accuracy is the proportion of both presences and absences correctly predicted. This measure is not without its problems, as there is some degree of accuracy that can occur by chance.

Table 4.1: The framework of a confusion matrix. The columns indicate presences (1) and absences (0) from a model's predictions, whereas the rows indicate presences (1) and absences (0) from actual observations. True positives (TP) are instances where the model correctly predicted presences, true negatives (TN) are instances where the model correctly predicted absences, false positives (FP) are instances where the model incorrectly predicted absences as presences and false negatives (FN) are instances where the model incorrectly predicted presences as absences.

		Model	
		1	**0**
Observations	**1**	TP	FN
	0	FP	TN

Kappa provides a measure of proportional accuracy (Pearson *et al*., 2004). It is an improvement on overall accuracy as it incorporates a measure of accuracy expected to occur by chance. It ranges between -1 and 1 where a value of 1 indicates perfect agreement and a value of 0 or less implies that the model is no better than random. The main criticism against kappa is that it is influenced by prevalence and is dependent on a threshold (Allouche *et al*., 2006; Pearson *et al*., 2004). The former makes it unsuitable for comparing among different studies or species (irrelevant to this study), while the latter does not allow for evaluation across the full spectrum of thresholds, but only evaluates on a binary level. It is however possible to calculate kappa across a range of thresholds and to find the optimum threshold at which maximum k is obtained. Allouche *et al.* (2006) proposed a measurement, TSS (previously known as the Hanssen-Kuipers discriminant), that improves on kappa. TSS compares the number of correct predictions minus those attributable to chance, to that of a hypothetical set of perfect predictions. Similar to kappa it

takes into account both omission and commission errors and chance, and is also presented on the same scale as kappa. TSS however is not affected by prevalence.

Table 4.2: Evaluation methods calculated using the entries of the confusion matrix, where $n = TP + FN + FP + TN$.

Method	Formula
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Overall accuracy	$\frac{TP + TN}{n}$
True Skills Statistic (TSS)	$\frac{(TP \times TN) - (FP \times FN)}{(TP + FN)(TN + FP)}$ sensitivity + specificity – 1
Cohen's kappa	$\frac{\left(\frac{TP + TN}{n}\right) - \frac{(TP + FP)(TP + FN) + (FN + TN)(TN + FP)}{n^2}}{1 - \frac{(TP + FP)(TP + FN) + (FN + TN)(TN + FP)}{n^2}}$

4.2 Methodology

4.2.1 Objectives

The two suitable habitat maps that were chosen from the different SDM building strategies (SDM1 from Section 3.3.1 and SDM2 from Section 3.3.2), joined with the available life-history parameters of the European Starling in South Africa (Table 2.1), were used to develop a dynamic IBM that simulated the starling's range expansion process in Southern Africa. Parameter sensitivity analysis was performed on various parameters to find an optimal model that best fit historical data of the range expansion process and best fit the current distribution of the starling. The former was evaluated with historical records by means of reduced major axes regression analyses, while the latter was evaluated with SABAP 1 records by means of AUC and methods derived from confusion matrix entries. Null models were used for studying several processes in isolation and examining whether the models performed better than random. The top performing models were ultimately used for making future projections of the starling's distribution in Southern Africa as well as for studying their distribution dynamics and dispersal routes.

Sensitivity analysis was essential in choosing the best fit model and to find the range of conditions appropriate for the study species where the model's predictions hold true (Higgins *et al*., 1996b). This is especially necessary when information regarding a species is uncertain or lacking and parameter estimates had to be made when developing the model. Carey (1996) studied the British distribution of *Himantoglossum hircinum*, the Lizard Orchid, and how climate change could affect their distribution. Carey applied a similar strategy in which the parameters were varied until an optimal result was found, after which he studied how the population would change with changing climate through varying survival probabilities.

4.2.2 Model Formulation

The scale at which my model was built had a resolution of 10 km x 10 km grids and an extent that covered most of South Africa, with 26˚S as the northern most latitudinal cut off line. Each time step was equivalent to a year and the model simulation began in 1897, with the release of 18 individuals in Cape Town (18.424˚E and 33.925˚S). 115 time steps were then simulated up to the year 2012. When referring to an individual I am in actual fact referring to a flock as the starlings tend to disperse in flocks (Fischl and Caccamise, 1985) and for simulation purposes it was much faster to simulate a group of individuals as one entity than each individual by itself. The simulation was consequently initiated using three flocks. A reproduction rate of 0.527 per year introduced new individuals into the population at each time step. The reproduction rate was calculated as a multiplication of the nest success, clutch size, juvenile survival rate and the proportion of females in the population (assumed to be half of the population), which resulted in $r = 0.74 \times 4.4 \times 0.324 \times 0.5 = 0.527$ (values from Table 2.1). The adults remained within the cell where they had nested while only the juveniles were given the opportunity to disperse and explore new habitats. This was consistent with observational field knowledge discussed in Section 2.1.2. A schematic illustration of the model is presented in Figure 4.1.

Each juvenile flock was given the option to choose between *n* number of different sites (*n* = 2, 5 or 10) when searching for a new nesting location. This decision represented the European Starling's cognitive ability to discern between the qualities of different habitat sites as well as the memory of places already visited. Learning is a product of different experiences, and the memory of these past experiences must be considered as an internal state of the organism (DeAngelis and Mooij, 2005). Their spatial memory is important for optimal foraging and by giving the starling this capacity I allowed them to choose the best suitable nesting site that would maximize their fitness in terms of both habitat quality and energy budget (flight distance).

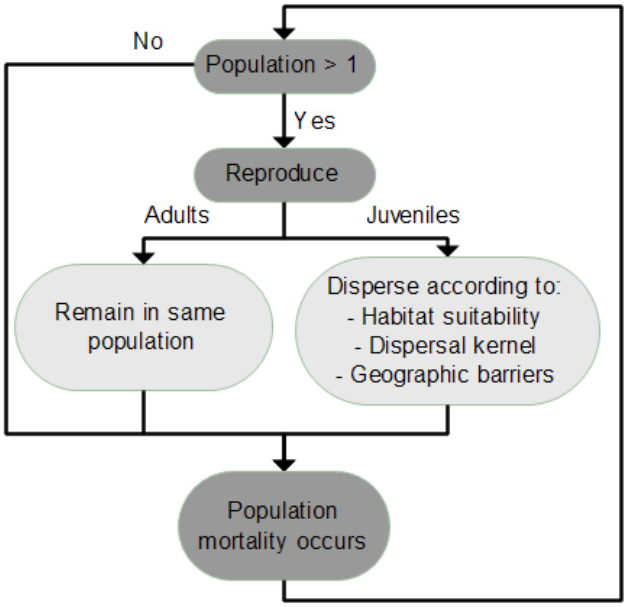


Figure 4.1: A simplified schematic illustration of my individual based model for the European Starling in Southern Africa. Dark grey boxes denote population processes and light grey boxes denote individual flock processes. Only juveniles were given the opportunity to disperse while adults remained at their nesting sites. Those individuals that survived after population mortality took place would form the new adult population in the following time step *t* +1.

Dispersal rates need careful consideration when modelling a species' range expansion. A dispersal kernel describes the dispersal ability of an individual, or group of individuals, as a function of distance. The various abilities and disabilities of the study species to disperse to a new suitable habitat site should be considered such that uncertainty in dispersal scenarios is reduced (Franklin, 2010). Approaching this concept, one can either apply a single estimate of dispersal rate by using a dispersal kernel, or one could impose constraints on their movement based on barriers, abundance across cells and other demographic factors such as fecundity and mortality. Including such factors increases the model's realism and significance.

Choosing a new location for each flock in my IBM was based on the following sequence of events:

(i) *Dispersal distance*

The dispersal kernel, from which the dispersal distance was obtained, was an inverse power law function with exponent -1.5 (from Section 2.2.4). A probability distribution function was calculated by integrating $\int_a^b cx^{-1.5}dx = 1$ and solving for the coefficient c (a and b are the minimum and maximum dispersal distances, respectively). Weights were attributed to every 1 km interval according to the integral of that interval. A random dispersal distance was then chosen from this probability distribution. Although the probability for a long distance dispersal event was very small, it existed nonetheless and therefore long distance dispersal events could

occur. The probability of dispersal with respect to the dispersal distance is shown in Figure 4.2, with maximum dispersal distances varying between 100 km, 200 km and 300 km.

(ii) *Dispersal angle*

A random angle between 0˚ and 360˚ was chosen.

(iii) *Spatial boundaries*

The new *x-y* coordinate was only considered when it fell within the spatial boundaries of the model, that is, within a grid and not outside the defined Southern African modelling extent. When the new coordinate fell outside these boundaries it was rejected and a new coordinate was generated.

(iv) *Geographic barriers*

An elevation barrier of x metres between two neighbouring grids restricted the flocks' movement (x varied between 100 m, 200 m and 300 m). Instead of crossing this barrier they remained in the grid adjacent to the barrier. The reasoning behind incorporating an elevation barrier was because of their preference to low lying areas and avoidance of mountainous areas (Link, 2004), also demonstrated in Figure 2.6 where a decrease in starling density is evident with increasing elevation. This parameter will further on be referred to as elevation access, as the birds have flight access to 100 m, 200 m or 300 m of elevation.

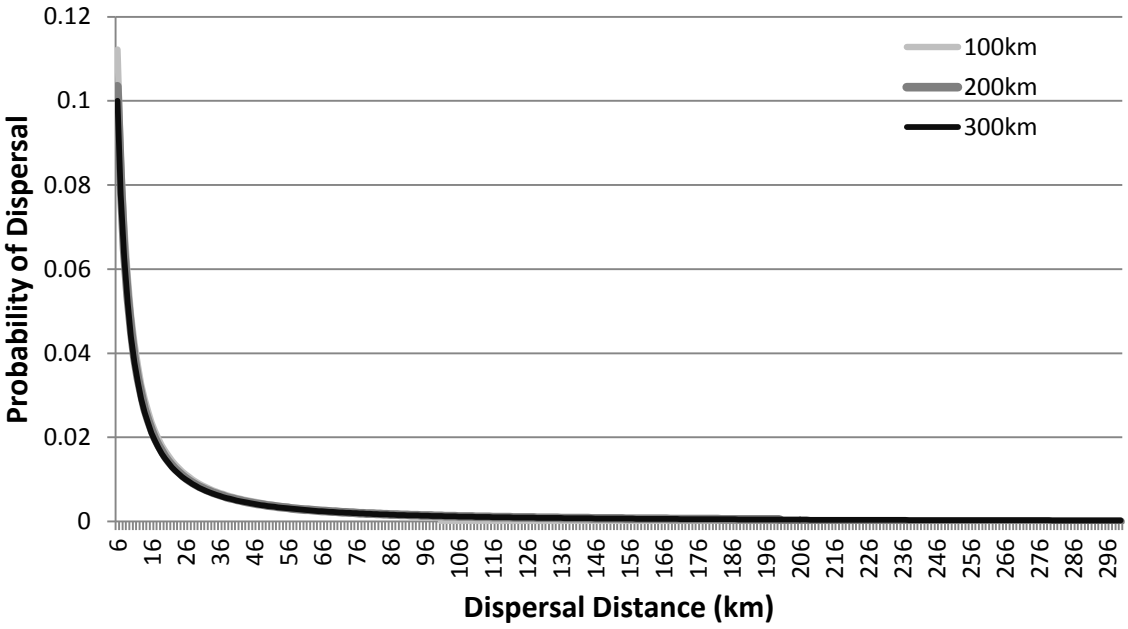


Figure 4.2: Dispersal kernels for varying scenarios of maximum dispersal distances from 5 km up to 100 km, 200 km or 300 km.

Once all the *n* different sites have been chosen according to these four steps, the sites were weighted against each other given their habitat suitability (HS) and dispersal distance from current location i to new location j (dis_{ij}) such that the site j for which $\left(\frac{1}{dis_{ij}}\right)\left(\frac{HS_j}{HS_i+\varepsilon}\right)$ was maximized (where $\varepsilon = 0.0001$) were chosen as the new location for the flock. I assumed that flocks within a cell were uniformly distributed and each flock was appointed the allocated suitable habitat and elevation values of the grid in which its *x*-*y* coordinates were located.

After each juvenile had the opportunity to disperse and new populations were established, population mortality within each grid took place depending on that grid i's habitat suitability (HS_i) and according to a negative exponential function $e^{-\lambda HS_i}$, where λ was calculated such that the mean of the function, hence the average mortality rate, was $\frac{1}{\lambda} = 0.353$ for the adults (from Table 2.1). The juveniles' mortality rate has already been taken into account within the reproduction function. 20 flocks were appointed as the ceiling number for each grid.

The elevation data (digital elevation map, DEM) was obtained from the CGIAR-CSI database (Consortium for Spatial Information of the Consultative Group for International Agricultural Research) (Jarvis *et al*., 2008) at approximately 90 m resolution (Figure 4.3). As with the predictor variables for my SDM, the elevation data and the habitat suitability maps (whose output was at 1 km resolution) were resampled to 10 km resolution using bilinear resampling (explained in Section 3.2.4.1). These data sets were edited such that they had the same extent (15.05˚E, 32.95˚E, 26.04˚S, 34.64˚S) and their grids were aligned accordingly. The corresponding elevation and habitat suitability values were extracted for each grid using *sample*, a spatial analyst tool in ArcGIS.

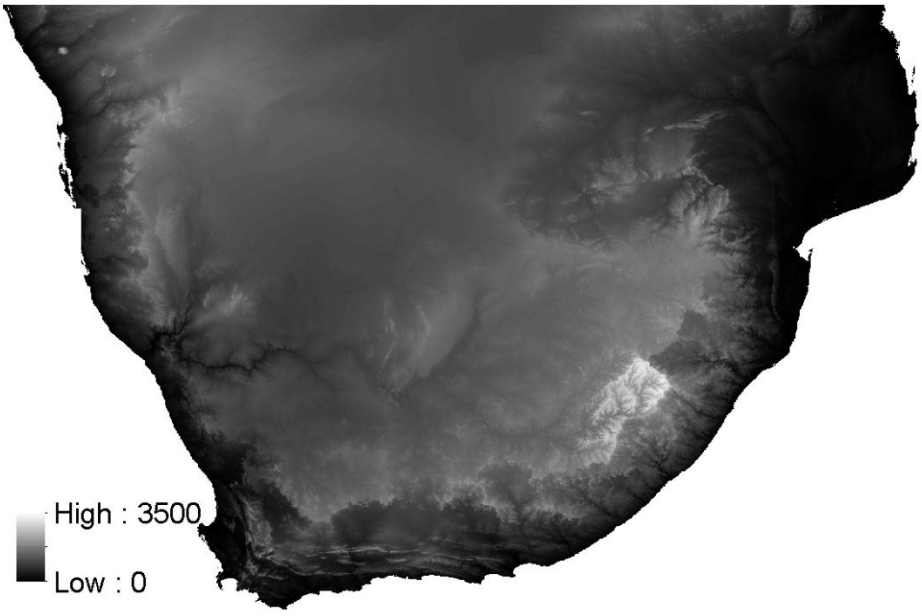


Figure 4.3: Digital elevation map for Southern Africa from CGIAR-CSI expressed in metres and displayed at 90 m resolution.

My parameter sensitivity analysis was based on varying the following parameters: the habitat suitability map, elevation access, number of dispersal choices, maximum dispersal distance and the decision function for choosing a new location (Table 4.3). This accumulated to 54 simulations for each habitat suitability map, and 108 simulations in total. The reason for doing a parameter sensitivity test was due to uncertainty regarding the influence of geographic barriers, dispersal distances and the degree of cognitive ability on the starling's decision making. Varying these parameters was necessary to find the best model fit and for studying the range of conditions appropriate for the starling's distribution in Southern Africa.

Table 4.3: Range of parameter values used for my model's parameter sensitivity analysis. For each simulation one option of each parameter was implemented, leading to a total of 108 combinations.

Habitat suitability map	SDM1 SDM2
Elevation access	100 m 200 m 300 m
Number of dispersal choices	2 5 10
Dispersal distance	5 km to 100 km 5 km to 200 km 5 km to 300 km
Decision function for choosing best new location	$\left(\frac{1}{dis_{ij}}\right)\left(\frac{HS_j}{HS_i+\varepsilon}\right)$ $\left(\frac{HS_j}{HS_i+\varepsilon}\right)$

My IBM was implemented in Mathematica 9.0 (Wolfram Research Inc). R (R Core Team, 2013) was also used, but was not as efficient for simulating multiple loops. The first attempt at modelling the above procedure was very time consuming. This was because each flock was tracked in great detail and the code was meticulously programmed in a manner which was inefficient for running multiple simulations. Those models took approximately two weeks to complete a single simulation that consisted of 115 time steps. An updated modelling approach was required and the code was improved accordingly. All the information for each grid was written in table format, including the coordinates, habitat suitability, elevation and number of adult and juvenile flocks per grid for each time step. On this table various rules were applied such as reproduction and mortality. Each juvenile flock was then extracted and modelled separately during the dispersal loop. This model completed a simulation in approximately two days; this was the final code I used for all subsequent model implementations.

4.2.3 Model Evaluation

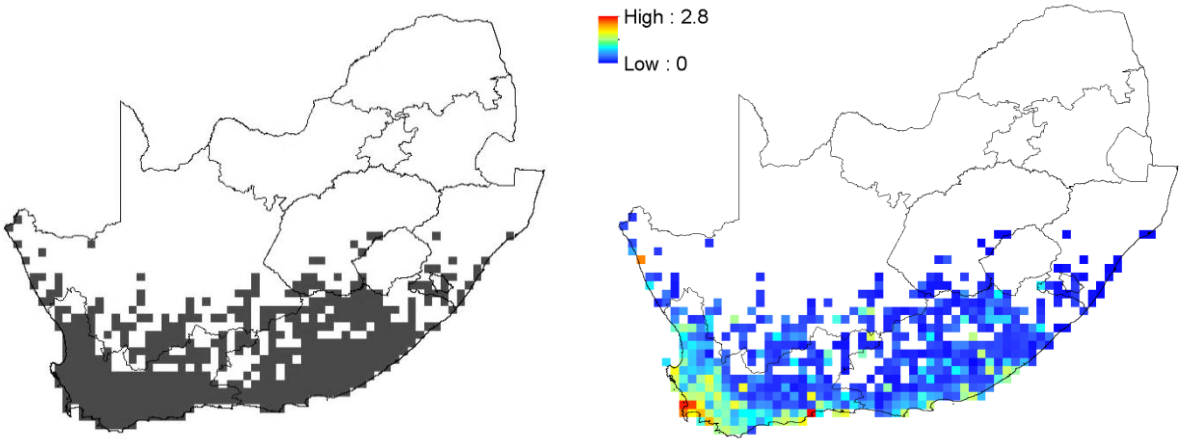
Model evaluation targeted two phases of the starling's range expansion, namely the invasion process and the current distribution. The former was evaluated with historical records, using linear regression analysis, while the latter was evaluated with SABAP 1 records using various confusion matrix measures and AUC. Null models were designed to study several processes in isolation.

4.2.3.1 Evaluation using Historical Records

Historical records (Harrison *et al*., 1997; Hui *et al*., 2012) provided me with information regarding specific locations and dates of the first recordings of the European Starling in Southern African towns (all records are in South Africa except for one record in Namibia). Figure 2.3 displayed a snapshot of this data, which in total consisted of 46 records shown in Figure 4.5 (i). The model simulations kept track of the starling occurrences for each grid over the 115 time steps. This enabled me to compare the time of the starling's arrival at those 46 locations according to my model as opposed to the actual observed time of arrival. For fitting model predictions with observations, ordinary least squares regression (OLS) is most often used and such linear model fitting was carried out in Mathematica. Reduced major axis regression (RMA) analysis was also performed. This analysis technique is an improvement on OLS regression as it can deal with skewed distributions of residues (Bohonak and Van der Linde, 2004; Hui *et al*., 2010). Hui *et al.* (2010) showed that RMA most closely matched the distribution of observed data when compared to OLS, adjusted and nonlinear methods. Regression coefficients, their standard deviations and the coefficient of determination (R^2) were calculated for both regression methods.

4.2.3.2 Evaluation using Current Records

An independent data set, SABAP 1, was used to evaluate my model's predicted distribution after 100 time steps which corresponded to the year 1997, the year when SABAP 1's data collection were completed. SABAP 1 was in the format of presence-absence data as well as relative density (Figure 4.4). My model's predicted distribution was converted to relative density and using similar calculations to that used for both SABAP 1 and SABAP 2 as described in Section 2.3.1. The relative density was in turn converted to presence-absence data by choosing an appropriate threshold value, which was 0 since this was the threshold value applied to SABAP 1.



(i) (ii)

Figure 4.4: European Starling's (i) presence-absence and (ii) relative density for the year 1997 (100 years since introduction) according to SABAP 1 data.

My model's distribution data were rescaled from a 10 km to a 25 km resolution as that was the resolution of the evaluation data set, SABAP 1. Rescaling was carried out by combining all 10 km grids that fell into the nearest 25 km grid. Confusion matrices were generated for all model simulations from which overall accuracy, sensitivity, specificity, TSS and kappa could be calculated. Additionally, AUC was also calculated for each simulation using the package *pROC* in R (Robin *et al*., 2012).

4.2.3.3 Null Models

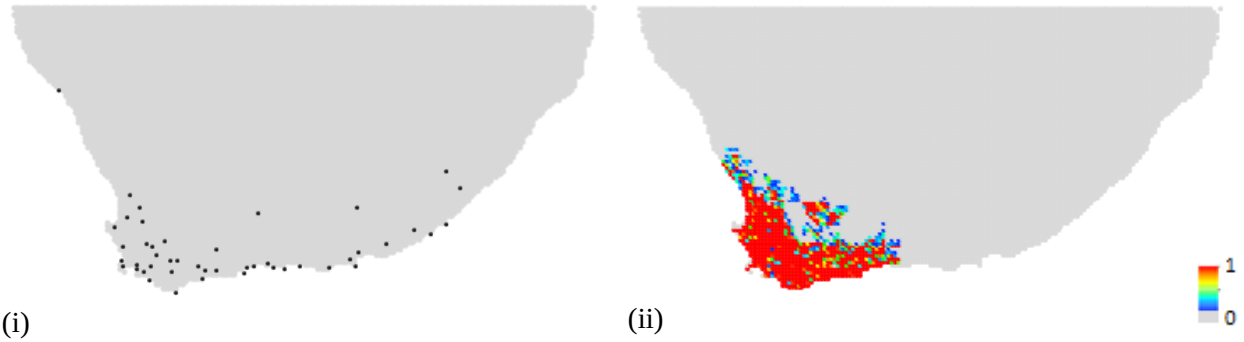
The reason behind developing null models was to study the importance of various processes by changing model sensitivity and complexity. An appropriate null model is one that retains everything in the original model design, but excludes only the factor(s) of interest. My first null model excluded elevation access while my second null model excluded the starling's cognitive ability, hence, no choices were given for a new location but random dispersal (according to the defined dispersal kernel) stipulated the new location (so no dispersal function was implemented). The final null model excluded both elevation and cognitive ability. The null models were evaluated against those models that retained the same parameter settings that each of them had respectively (for example the same dispersal distance, same habitat suitability map, etc.).

4.3 Results and Discussion

The corresponding names and descriptions for all 108 model simulations are given in Appendix B. From here on the simulations will be referred to by those names. Numerous evaluation criteria were used as discussed in Section 4.2.3. Choosing the model that performed best was challenging as the criteria differed in the various model properties they examined. For historical records' evaluations I decided on studying only R^2 and the slope as calculated by RMA. The y-intercept and standard error are inherently dependent on those two criteria so it was unnecessary to report them as well. RMA and OLS gave similar results, but RMA was preferred above OLS due to its ability to take into account skewed distributions of residues. All the historical records' evaluation results are presented in Appendix C and Appendix D, in this section I only commented on R^2 and the RMA slope. For the current records' evaluations I decided on studying overall accuracy, kappa, TSS and AUC as sensitivity and specificity only measure either the proportion of correctly predicted presences or absences and were already incorporated in some of the other criteria (all these results are presented in Appendix E). The models were ranked according to each criterion and the sum of rankings was used as an overall indicator of performance for historical and current records' evaluations, separately as well as combined (presented in Appendix F). The relative density plots presented in this section were normalized to a scale between 0 and 1.

4.3.1 Evaluation using Historical Records

I compared the year that the European Starling was recorded in each of the 46 locations from the historical records (Figure 4.5 (i)), to the year that the model first predicted their arrival at those locations. Some of the model simulations were very restrictive in that even after 115 time steps the starling flocks were not able to reach all 46 locations (Figure 4.5 (ii)). Since my simulations only consisted of 115 time steps (1897 – 2012), I did not extrapolate when the starlings would get to those locations. Instead I appointed a pseudo prediction of the year 2013 to all such locations that were not occupied by 2012.



(i) (ii)

Figure 4.5: (i) The 46 locations from historical records used for model evaluation and (ii) the European Starling's relative density in a restrictive example of a model simulation after 115 time steps (Sim1.18). Some of the 46 locations were not occupied by the European Starling after 115 time steps. Such locations were appointed the year 2013 (116th time step) as a pseudo prediction in order to perform and compare model evaluations.

The European Starling's range expansion can be separated into a two-phase range expansion according to the historical records (Hui *et al.*, 2012) (Figure 4.6). Initial range expansion occurred slowly as the starlings took roughly 30 to 60 years to move 200 km from their introduced point, Cape Town. After this initial phase they took only 40 to 50 years to move a further 1000 km, representing a faster second expansion phase. An appropriate model that will perform well when evaluating against the historical records is one that can simulate these slow and fast expansion phases separately.

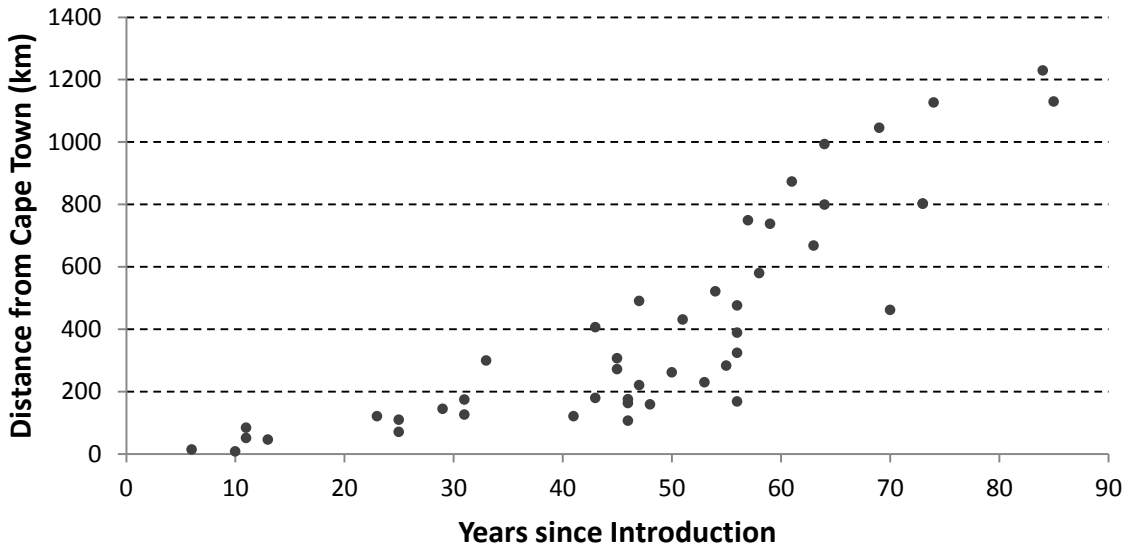


Figure 4.6: The two-phase range expansion of the European Starling in Southern Africa as depicted by the 46 historical records, illustrating the relationship between the years (since introduction) that they were first recorded at those locations and the distance away from Cape Town (the point of introduction). An initial slow range expansion is observed followed by a faster second range expansion.

According to the summed rankings of R^2 and the absolute difference of the RMA slope to 1, Sim1.41 performed the best of all 108 models (Table 4.4). Sim1.41 had an elevation access of 300 m, the flocks were able to disperse up to 300 km per year, only two dispersal choices were given and new habitat was chosen based only on habitat suitability (HS). This is in fact the least restrictive model of all simulations as the flocks were given maximum dispersal distance, maximum elevation access with minimal cognitive ability allowing for more random movement than movement limited by a complicated foraging process. This non-restrictive model therefore lead to a faster range expansion (Figure 4.7) when compared to the other simulations.

Table 4.4: The three best and three worst rated models according to the summed rankings of their RMA slope and R^2 when compared with historical records. Shown here are the selected evaluation criteria of slope as depicted by RMA, the coefficient of determination, R^2, and the models' rankings according to these criteria. The full table of results is presented in Appendix C.

	Historical Records' Evaluations			Rankings		
Model	Slope	Absolute Difference of Slope to 1	R^2	Slope	R^2	Summed Rankings
Sim1.41	1.051	0.051	0.699	6	7	1
Sim2.24	0.985	0.015	0.681	2	18	2
Sim2.39	0.982	0.018	0.683	3	17	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sim1.36	0.460	0.540	0.588	99	95	106
Sim1.18	0.460	0.540	0.566	100	99	107
Sim2.52	0.445	0.555	0.548	107	101	108

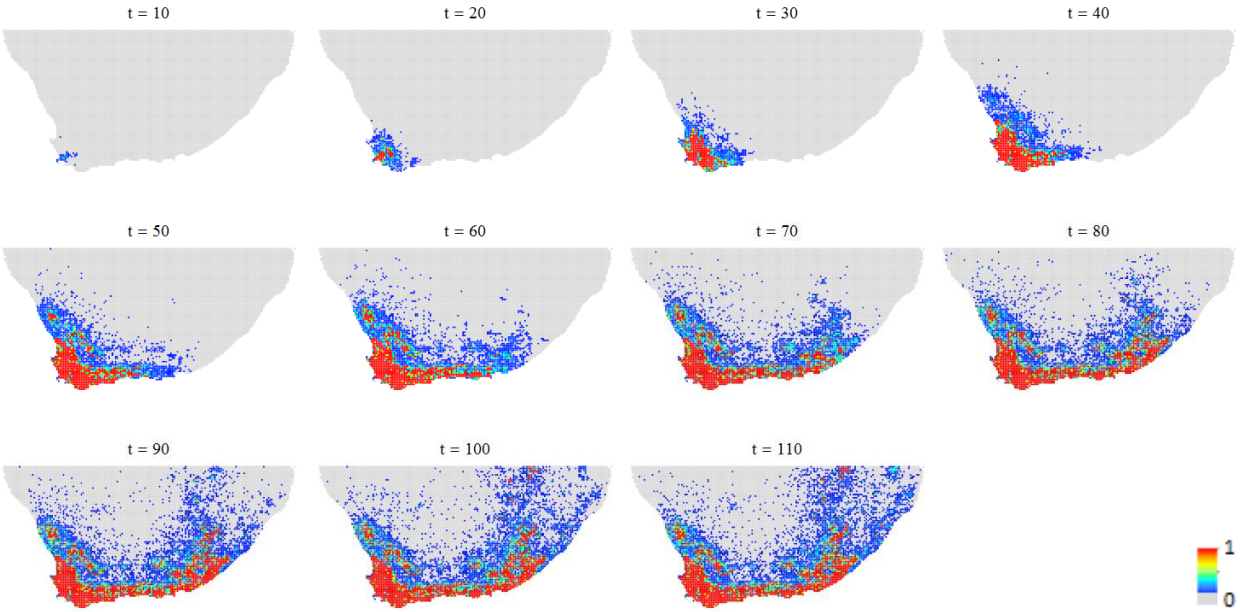


Figure 4.7: The range expansion of the European Starling (relative density in Southern Africa per 10 km grid over 110 time steps) as modelled by Sim1.41 (SDM1, DEM300, CH2, 5 to 300km, HS), the best performing model based on summed rankings of historical records' evaluations.

All three top models gave similar evaluation results ($R^2 > 0.68$, slope ≈ 1) and were also similar in their parameter values. They had good elevation access (200 m or 300 m), only two dispersal choices, and dispersal kernels with great dispersal distances (maximum distance of 200 km or 300 km). All of these parameter settings contributed to less restrictive models. Figure 4.8 illustrates the fast expansions of these three less restrictive models as they predicted the starling's occurrences in the 46 towns on average faster than the observations (most points lie to the right of the dashed line of perfect fit).

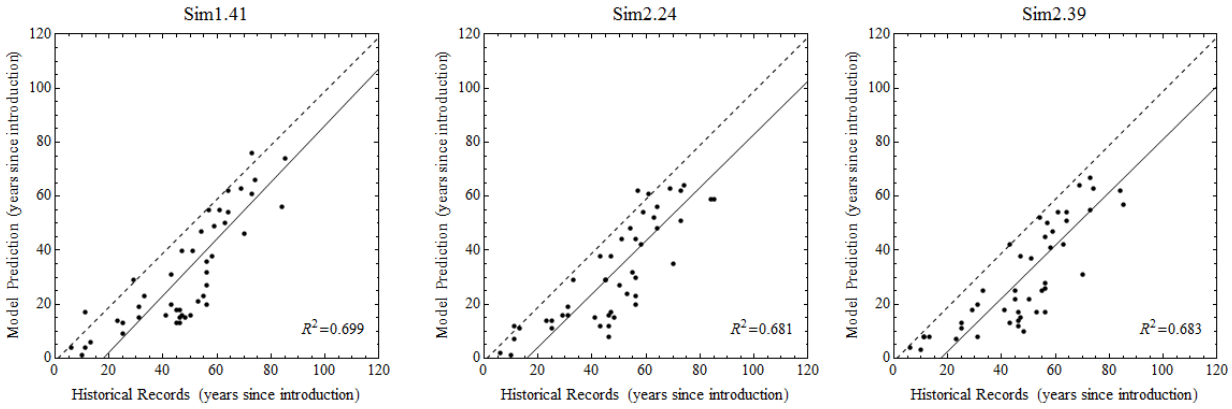


Figure 4.8: Regression analyses for the top three models based on summed rankings of historical records' evaluations. The models and their parameters were from left to right: Sim1.41 (SDM1, DEM300, CH2, 5 to 300km, HS), Sim2.24 (SDM2, DEM200, CH2, 5 to 300km, HS DIS) and Sim2.39 (SDM2, DEM300, CH2, 5 to 200km, HS). The dashed lines represent a perfect fit. All the 108 regression plots can be viewed in Appendix D.

The three models performing worst were Sim1.36 (DEM200, CH10, 5 to 300km, HS DIS), Sim1.18 (DEM100, CH10, 5 to 300km, HS DIS) and Sim2.52 (DEM300, CH10, 5 to 200km, HS DIS). All three had ten dispersal choices and a dispersal function incorporating both HS and distance when choosing a new location, which placed restriction on the flocks' movement through increasing their cognitive ability. The elevation access and dispersal distance parameters were inconsistent among these three models. It therefore seems that cognitive ability is a dominating factor in predicting how well the models perform when evaluating with historical records, suggesting that simple comparisons of habitat sites are more realistic than complicated ones.

Most of my models lacked the ability to predict the slow expansion of the first phase of range expansion and fast expansion of the second phase. This could be due to an error in the historical records, an error in model setup, or a combination of both. A general trend can be seen among the majority of simulations (Figure 4.9 and Appendix D) where the model predicted a faster range expansion than the records during the initial ±50 years whereas afterwards the model predicted a slower range expansion than the records (slope > 1).

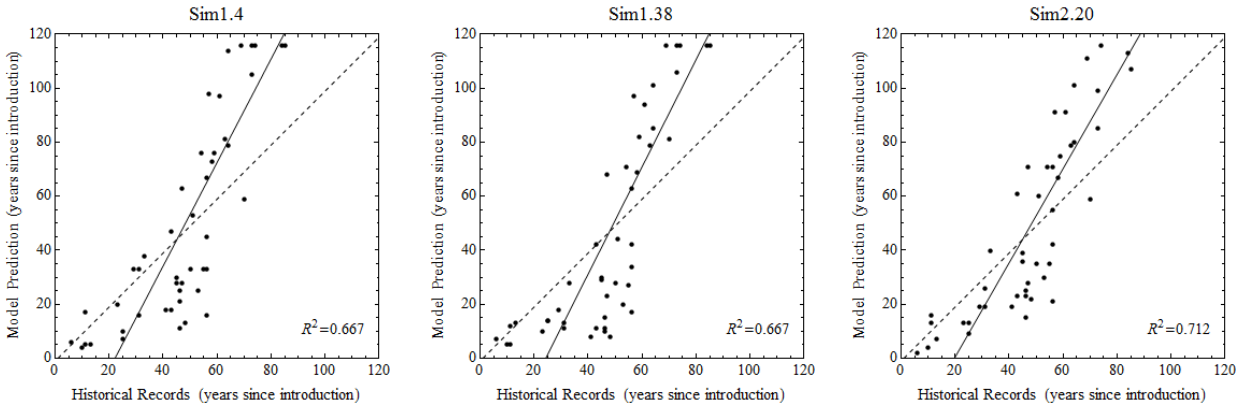


Figure 4.9: Regression analyses for a random selection of models, representative of the majority of the models (Appendix D), illustrating the general trend in model performance when evaluating with historical records. A faster expansion of the model against the records in the first expansion phase and a slower expansion of the model against the records in the second expansion phase is observed. The dashed lines represent a perfect fit. The models shown here from left to right were Sim1.4 (SDM1, DEM100, CH2, 5 to 200km, HS DIS), Sim1.38 (SDM1, DEM300, CH2, 5 to 100km, HS DIS) and Sim2.20 (SDM2, DEM200, CH2, 5 to 100km, HS DIS).

I do not however expect a perfect relationship (dashed line in regression plots), mainly due to imperfect data collection. In Figure 4.10 we observe that certain data points are out of place with surrounding points concerning the year starlings were first recorded there and do not follow on each other chronologically across space (green and red highlighted records). Due to such anomalies I preferred using RMA over OLS as it took error in the observed variable into consideration. Despite this the interpretation of such results should still be done with caution.

Close to Cape Town the difference in years of first sightings close to each other is relatively large, amplifying the observation of a slow range expansion during the first phase. Three records stand out among the green highlighted records in Figure 4.10, they are the highlighted 41, 46 and 56 closest to Cape Town. According to the data it took the starlings up to 41, 46 and 56 years to travel only 120 km, 106 km and 167 km from Cape Town to these towns, respectively. These are especially low spread rates when keeping in mind the vast distances the European Starling tends to travel as mentioned in Section 2.2.4. Possible reasons could be that they did not travel very far in the initial years of population growth, going through an establishment phase and waiting for flocks to get bigger before going off on foraging expeditions, the environmental and/or geographical barriers were too intimidating, or the data collection at those points are misleading and the starlings were in fact present at those locations earlier on but were only recorded there in later years. In all the models' regression plots we observe at least one cluster of data points between 40 to 60 time steps that are over predicted by the models (Figure 4.8, Figure 4.9, Figure 4.11 and Appendix D), these are some of the green records in Figure 4.10.

The opposite is seen the further away you go from Cape Town, where a small difference in years of adjacent records amplify the fast range expansion of the second phase. First sighting records highlighted in red (Figure 4.10) stand out above the rest, as they appear at an earlier time than that of neighbouring

locations located closer to Cape Town. For a model to perform well according to the historic records, it either had to (i) simulate a very slow range expansion during the initial years, but thereby run the risk of under predicting the rest of the range expansion as well (Figure 4.11), and/or (ii) simulate a very fast range expansion during the second phase but run the risk of over predicting the initial phase (Figure 4.8 and Figure 4.9).

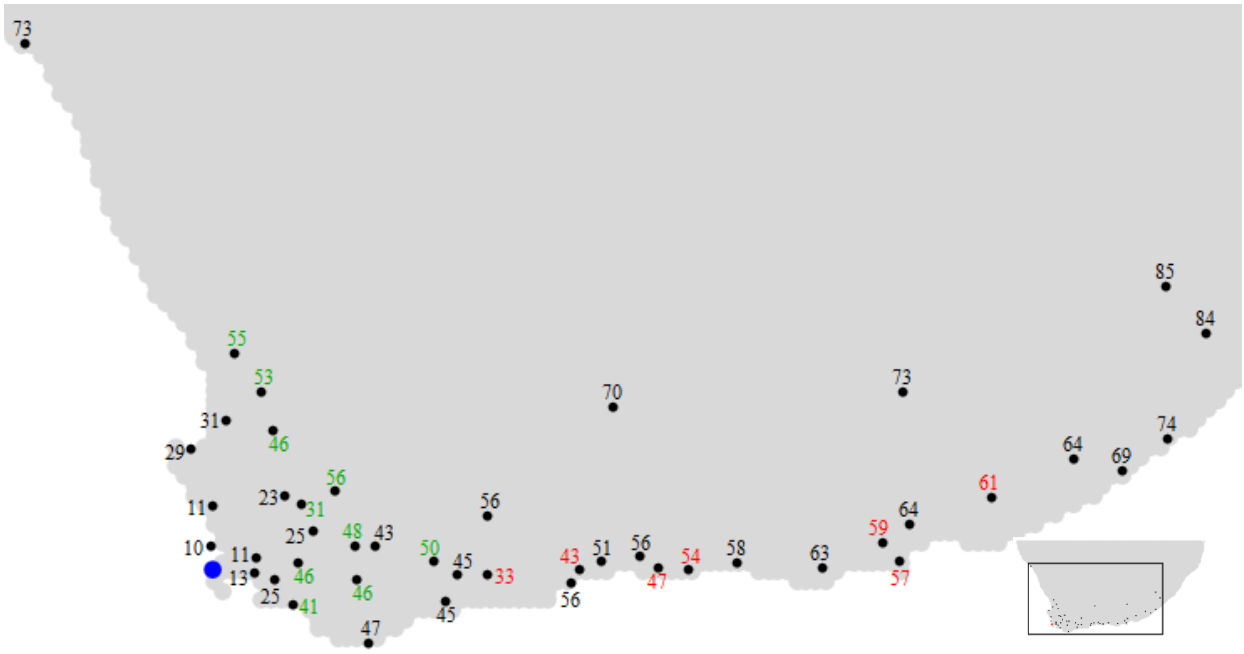


Figure 4.10: Historical records and the corresponding years since introduction that the European Starling was first sighted and recorded at various locations around South Africa and Namibia. The enlarged blue record is Cape Town, the point of introduction. Records highlighted in green represent observations suggestive of a slow first phase range expansion. Records highlighted in red represent observations suggestive of a fast second phase range expansion. Some of these records do not follow on each other chronologically across space.

Figure 4.11 shows a few examples that came close to predicting the initial phase of slow expansion. As with many other models, they were unable to predict the second phase's fast expansion. Additionally, they were also unable to simulate expansion up to some locations (evident in the group of pseudo predictions at the final time step, 116). All three these simulations had in common the number of choices (CH10), a minimal dispersal range (100 km or 200 km) and a decision function incorporating both HS and distance to new location. All of these parameter settings restricted the spread of the individuals despite the varying elevation barrier, thereby naturally forcing a slow expansion throughout the 115 time steps.

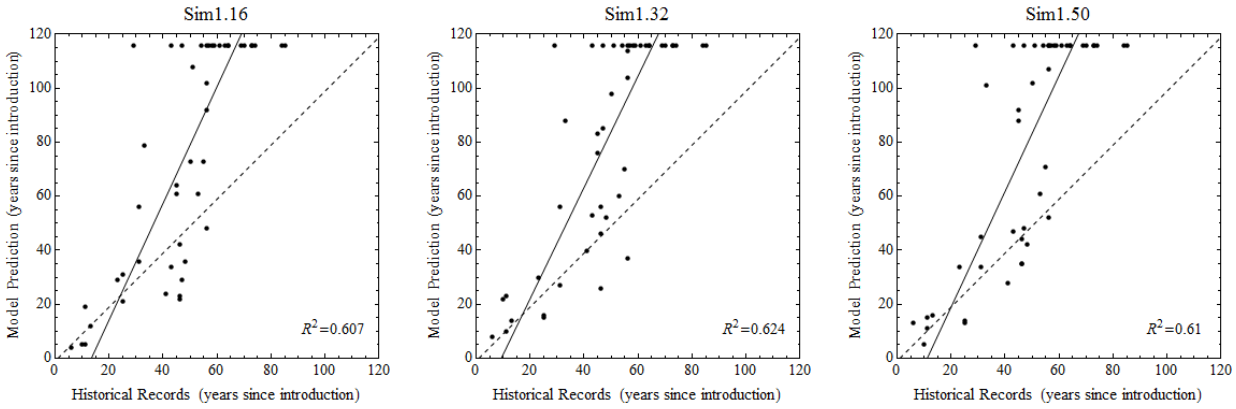


Figure 4.11: Regression analyses for a selection of models that best depicted the initial slow expansion phase. The models and their parameters were, from left to right: Sim1.16 (DEM100, CH10, 5 to 200km, HS DIS), Sim1.32 (DEM200, CH10, 5 to 100km, HS DIS) and Sim1.50 (DEM300, CH10, 5 to 100km, HS DIS). The dashed lines represent a perfect fit.

It is evidently very challenging to find a model that is able to predict the two-phase range expansion according to the evaluation criteria of the historical records. A final evaluation was conducted: finding the sum of squares between the data points and the perfect linear fit (dashed line). The three models that minimised the sums of squares are displayed in Figure 4.12 and the range expansion of Sim2.38, the model performing best according to this evaluation method, is displayed in Figure 4.13.

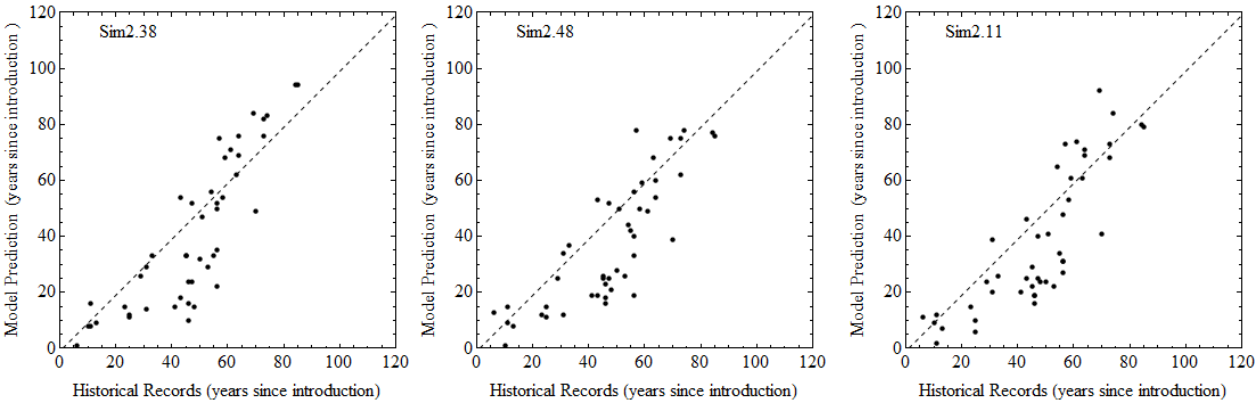


Figure 4.12: The three models that minimized the sums of squares between data points and the perfect linear fit (dashed line). The models and their parameters were, from left to right: Sim2.38 (SDM2, DEM300, CH2, 5 to 100km, HS DIS), Sim2.48 (SDM2, DEM300, CH5, 5 to 300km, HS DIS) and Sim2.11 (SDM2, DEM100, CH5, 5 to 300km, HS).

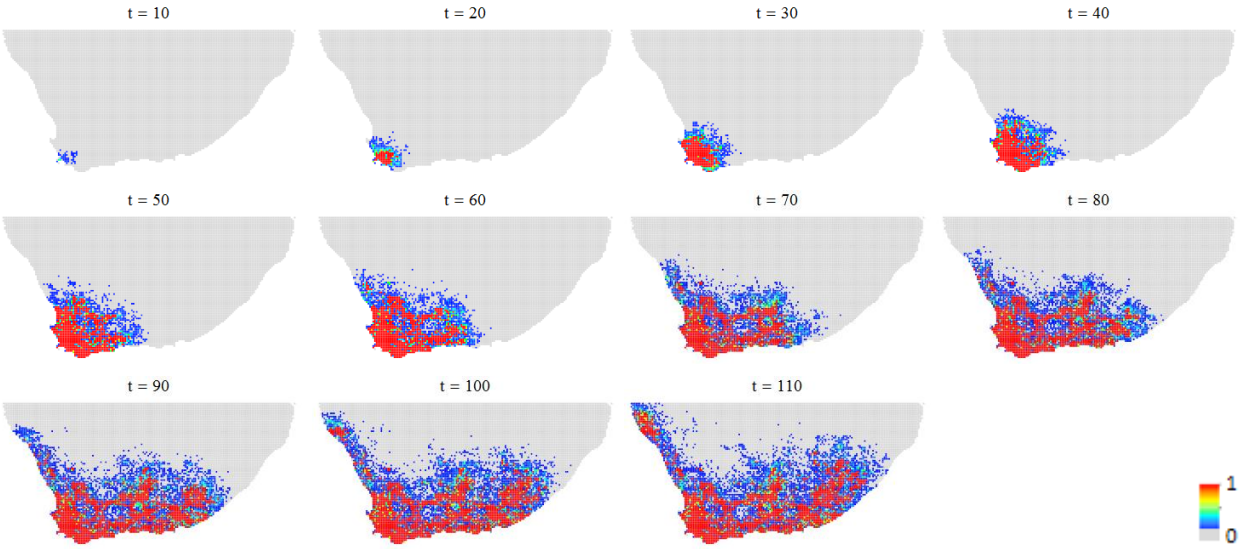


Figure 4.13: Range expansion of the European Starling as modelled by Sim2.38 (DEM300, CH2, 5 to 100km, HS DIS), the model that predicted minimal sums of squares between the data points and the perfect linear fit.

4.3.2 Evaluation using Current Records

A model that performed well when evaluating with current records had to have a distribution similar to that of SABAP 1 after 100 time steps. Sim1.6 was most successful in achieving such a distribution and was ranked highest according to the sum of all four evaluation criteria rankings (Table 4.5). Sim1.6 allowed the starling flocks flight access to only 100 m elevation and new habitat was chosen based on HS and distance, both of which greatly restricted the starling's long distance dispersal ability. This was despite the flocks having a dispersal kernel with a maximum dispersal distance of 300 km per year and allowing them to choose among two sites only.

Table 4.5: The three best and three worst ranked models according to summed rankings of current records' evaluation criteria, as well as best ranked models according to each evaluation criteria. All model results are shown in Appendix E.

	Current Records' Evaluations				Rankings				
Model	Overall Accuracy	Kappa	TSS	AUC	Overall Accuracy	Kappa	TSS	AUC	Summed Rankings
Sim1.6	0.817	0.593	0.664	0.882	9.5	6	1	10	1
Sim1.29	0.844	0.626	0.651	0.852	4	2	2	20	2
Sim1.45	0.830	0.796	0.844	0.602	5	5	3	18	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sim1.11	0.858	0.640	0.630	0.828	1	1	8	28	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sim1.39	0.726	0.461	0.585	0.906	75	37	15	1	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sim2.41	0.385	0.085	0.142	0.824	107	107	107	41	106
Sim2.42	0.384	0.084	0.141	0.826	108	108	108	38.5	107
Sim1.14	0.772	0.247	0.187	0.593	66.5	91	103	108	108

Sim1.41, which performed best in previous evaluations (Section 4.3.1), also had two dispersal choices and a maximum dispersal distance of 300 km, but had an elevation access of 300 m and only chose sites based on HS. The parameters that evidently dominated the different performances based on evaluation technique were elevation access and the dispersal function. As opposed to evaluating with historical records that required a relatively faster range expansion (Sim1.41, Figure 4.7), evaluating using SABAP 1 data required more restrictions on dispersal that prevented the flocks from expanding their range further than that of SABAP 1 within 100 years (Sim1.6, Figure 4.14).

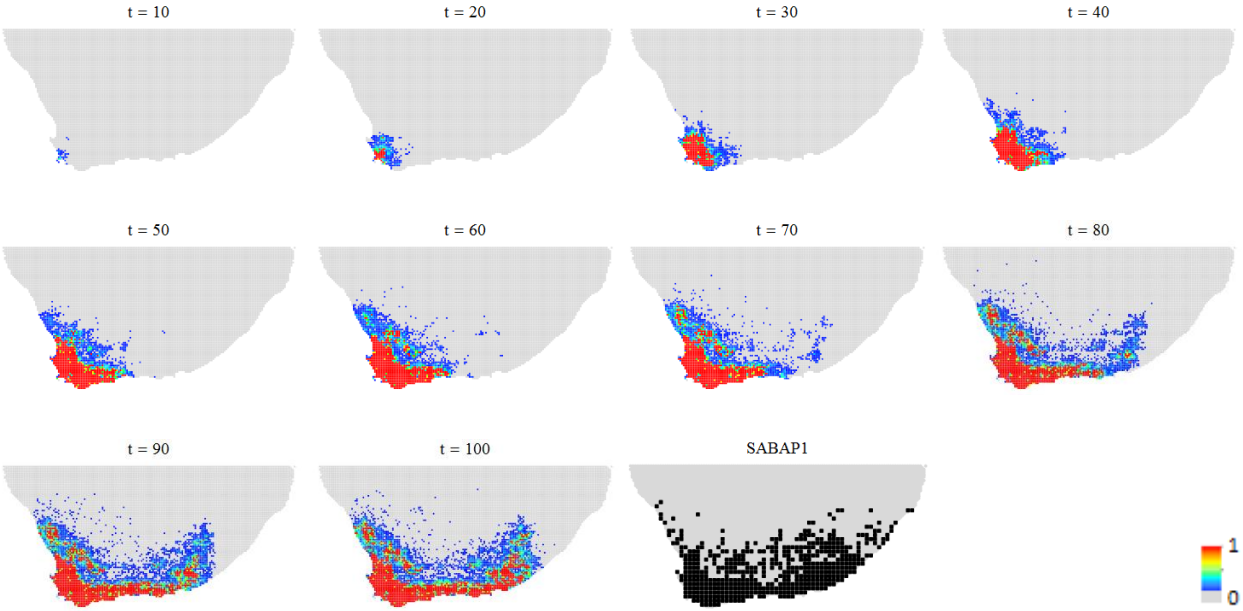


Figure 4.14: The range expansion of the European Starling as modelled by the best ranked model of current records' evaluation criteria, Sim1.6 (SDM1, DEM100, CH2, 5 to 300km, HS DIS), depicted as European Starling's relative density. The last map shows SABAP 1 distribution in 1997 equivalent in years to 100 simulated time steps.

When comparing the top three performing models to SABAP 1 (Figure 4.15), Sim1.6 lacked the ability to further predict occurrences along the eastern coastline where the starling was present in SABAP 1, but most accurately predicted occurrences inland when compared to the other two models. Distributions after 100 time steps as predicted by Sim1.29 and Sim1.45 appeared very similar to each other, but different to that of Sim1.6. These differences and similarities are due to the degree of cognitive ability in the models. Both Sim1.29 and Sim1.45 allowed the flocks' five dispersal choices and the decision of a new location was based only on HS. Their distributions were therefore highly related and dependent on HS (SDM1). Sim1.6 on the other hand allowed only two dispersal choices, while the dispersal function was based on both HS and distance. Fewer dispersal choices lead to greater randomness in dispersal routes, while both parameters combined lead to less dependence on HS and thus a more uniform distribution was obtained.

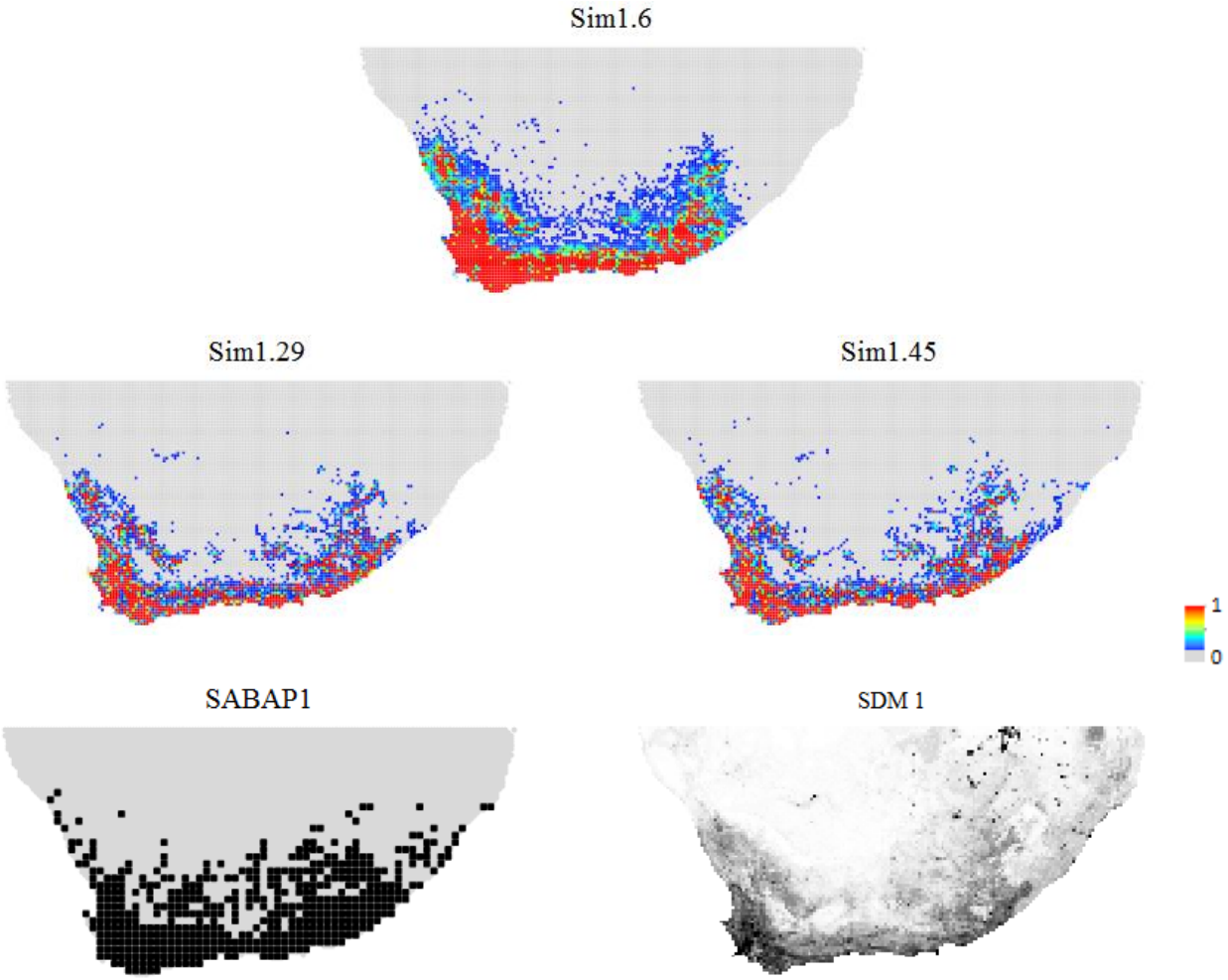


Figure 4.15: The relative density of the European Starling in 1997 as modelled by the top three models: Sim1.6 (SDM1, DEM100, CH2, 5 to 300km, HS DIS), Sim1.29 (SDM1, DEM200, CH5, 5 to 300km, HS) and Sim1.45 (SDM1, DEM300, CH5, 5 to 200km, HS). Also shown are SABAP 1 presences in 1997 and the habitat suitability map SDM1 that was employed in all three these models.

The three worst performing models were based on either the least restrictive (Sim2.41: 300 m elevation access, two dispersal choices, 300 km maximum dispersal distance, dispersal function based only on HS) or most restrictive (Sim1.14: 100 m elevation access, ten dispersal choices, 100 km maximum dispersal distance and dispersal function based on HS and distance) parameter settings, leading to distributions furthest away from that of SABAP 1 at both extremes (Figure 4.16). Interestingly, both Sim2.41 and Sim2.42 predicted very high AUC values (0.824 and 0.826, respectively) but low overall accuracies (0.385 and 0.384, respectively), while Sim1.14 predicted low AUC (0.593), but high overall accuracy (0.772), implying that both AUC and overall accuracy are highly dependent on sensitivity and specificity respectively.

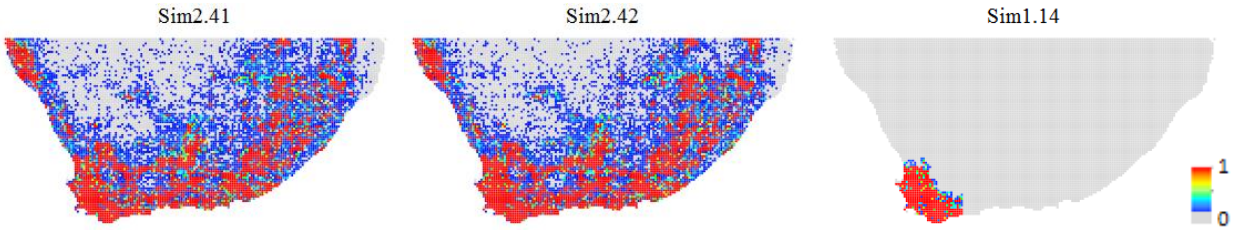


Figure 4.16: The relative density of the European Starling in 1997 as modelled by the worst three models, from left to right: Sim2.41 (SDM2, DEM300, CH2, 5 to 300km, HS), Sim2.42 (SDM2, DEM300, CH2, 5 to 300km, HS DIS) and Sim1.14 (SDM1, DEM100, CH10, 5 to 100km, HS DIS).

Out of all 108 models, Sim1.39 had the highest AUC (0.906), a sensitivity of 0.942 and specificity of 0.643, confirming my inclination that AUC is biased towards the proportion of correctly predicted presences. Sim1.39 indeed over predicted the SABAP 1 distribution (Figure 4.17) as most presences were accounted for, but many absences were predicted as presences. This is in contrast to the more balanced Sim1.6 that predicted a sensitivity of 0.866 and specificity of 0.798. In fact, Sim1.6 predicted the highest TSS (0.664) and performed best according to summed rankings, strengthening my initial motivation from literature (Section 4.1.1, (Allouche *et al*., 2006)) that TSS is a good method for evaluating binary data. TSS corrected for AUC's bias by taking into consideration both specificity and sensitivity. Kappa is a function of overall accuracy, illustrated by Sim1.11 which performed best with overall accuracy evaluations (0.858) as well as kappa evaluations (0.64). Manel *et al.* (2001) found kappa to be a more robust indicator of model performance as opposed to ROC measurements, which are strongly dependent on prevalence.

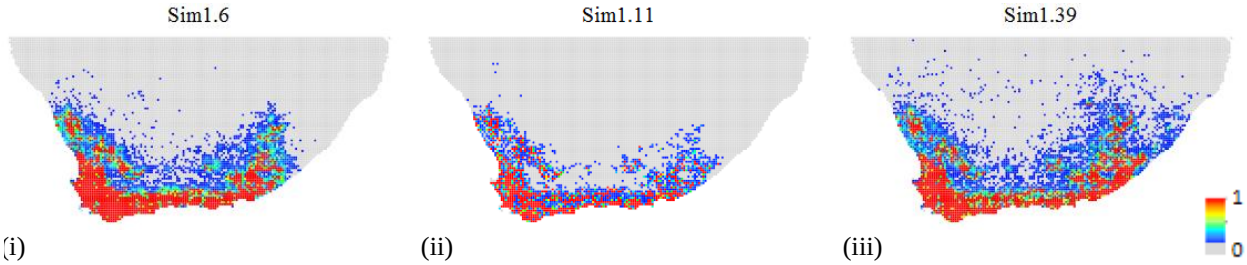


Figure 4.17: The relative density of the European Starling in 1997 as modelled by the models that performed best in (i) overall performance and TSS: Sim1.6 (SDM1, DEM100, CH2, 5 to 300km, HS DIS), (ii) overall accuracy and kappa: Sim1.11 (SDM1, DEM100, CH5, 5 to 300km, HS) and (iii) AUC: Sim1.39 (DEM300, CH2, 5 to 200km, HS).

All grids across the Southern African modelling extent were used for constructing confusion matrices. It might seem that this would automatically ensure a high proportion of true negatives which leads to a high value for specificity. However, since some models predicted that the starlings were indeed able to reach

grids across the entire modelling extent within 100 years (for example Sim2.41 and Sim2.42 in Figure 4.16), it was reasonable to use the entire extent for evaluations.

4.3.3 Evaluation using Historical and Current Records

I studied two historical and four current records' evaluation criteria (slope and R^2; overall accuracy, kappa, TSS and AUC). Measuring model performance using all these six criteria is biased towards current records' performances, so only kappa and TSS were used to balance the two evaluation methods. Kappa and TSS were chosen above overall accuracy and AUC due to literature support of these two methods being least biased (Allouche *et al*., 2006; Lobo *et al*., 2008; Manel *et al*., 2001). Continuing my evaluation technique, the models were ranked according to performance in each criterion and summed rankings were calculated. The top three models' results are shown in Table 4.6 and Figure 4.18. These were also the same three models that performed best even when all six evaluation criteria were used, strengthening my choice of choosing kappa and TSS above other evaluation criteria.

Table 4.6: The three best and three worst rated models according to summed rankings of two historical and two current records' evaluation criteria (Slope, R^2, kappa and TSS). All model results are shown in Appendix F.

	Historical Records' Evaluations			Current Records' Evaluations				Rankings		
Model	Slope	Absolute Difference of Slope to 1	R^2	Overall Accuracy	Kappa	TSS	AUC	Historical Records' Evaluations	Current Records' Evaluations	Combined Evaluations (Slope, R^2, Kappa & TSS)
Sim1.45	1.428	0.428	0.677	0.830	0.602	0.639	0.854	12.5	3	1
Sim1.11	1.680	0.680	0.691	0.858	0.640	0.630	0.828	15	6	2
Sim1.37	1.703	0.703	0.699	0.824	0.584	0.615	0.838	12.5	7	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sim1.50	2.144	1.144	0.610	0.780	0.284	0.217	0.608	100	102	106
Sim1.16	2.162	1.162	0.607	0.783	0.297	0.228	0.614	102	100	107
Sim1.14	2.089	1.089	0.602	0.772	0.247	0.187	0.593	92.5	108	108

For a model to perform well according to both evaluation methods it had to be able to show rapid expansion to imitate the fast second phase expansion of historical records while at the same time restricting the expansion to reach a distribution similar to that of SABAP 1. Sim1.45 was able to achieve this. It was ranked third according to the current records' evaluations and 12th according to the historical records' evaluations. Sim1.45 had five dispersal choices and a maximum dispersal distance of 200 km, which aided in restricting the starlings' distribution to the habitat suitability map SDM1, allowing for a good comparison with SABAP 1 distribution. At the same time it had 300 m elevation access and chose locations based on HS alone, which were all parameter settings that lifted dispersal restrictions and allowed for a good regression fit to the historical records (Figure 4.18). The models that performed worst had highly restrictive parameter settings as observed by their distributions after 100 time steps

(Figure 4.19). The predicted population growth and range size expansion of the European Starling across time for the three top performing models are shown in Figure 4.20.

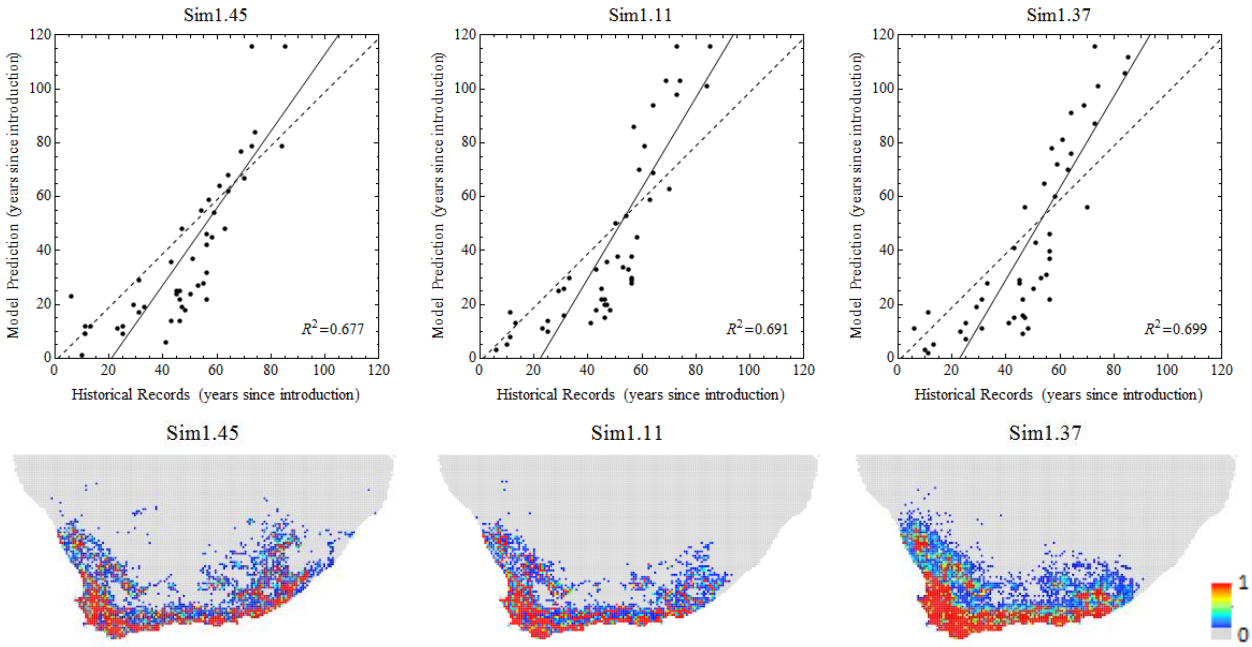


Figure 4.18: Regression analyses and relative densities (in 1997) of the European Starling as modelled by the top three models. These models and their parameter settings were, from left to right: Sim1.45 (SDM1, DEM300, CH5, 5 to 200km, HS), Sim1.11 (SDM1, DEM100, CH5, 5 to 300km, HS) and Sim1.37 (SDM1, DEM300, CH2, 5 to 100km, HS).

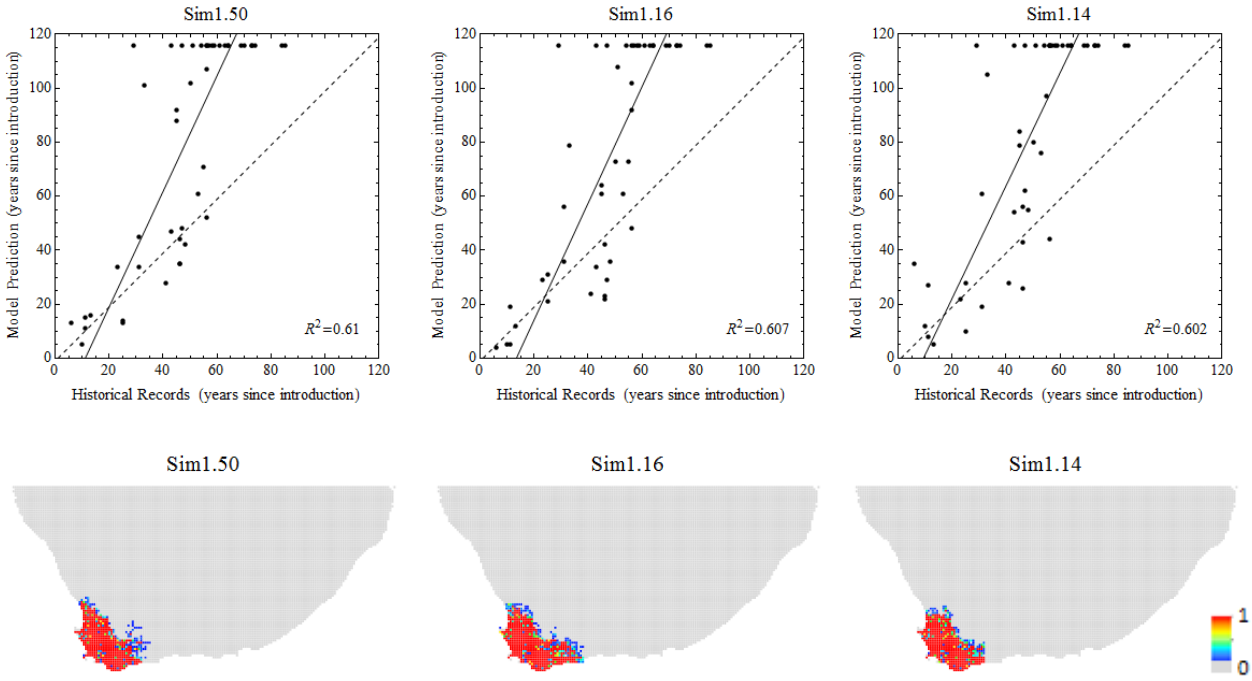


Figure 4.19: Regression analyses and relative densities (in 1997) of the European Starling as modelled by the three worst models. These models and their parameter settings were, from left to right: Sim1.50 (SDM1, DEM300, CH10, 5 to 100km, HS DIS), Sim1.16 (SDM1, DEM100, CH10, 5 to 200km, HS DIS) and Sim1.14 (SDM1, DEM100, CH10, 5 to 100km, HS DIS).

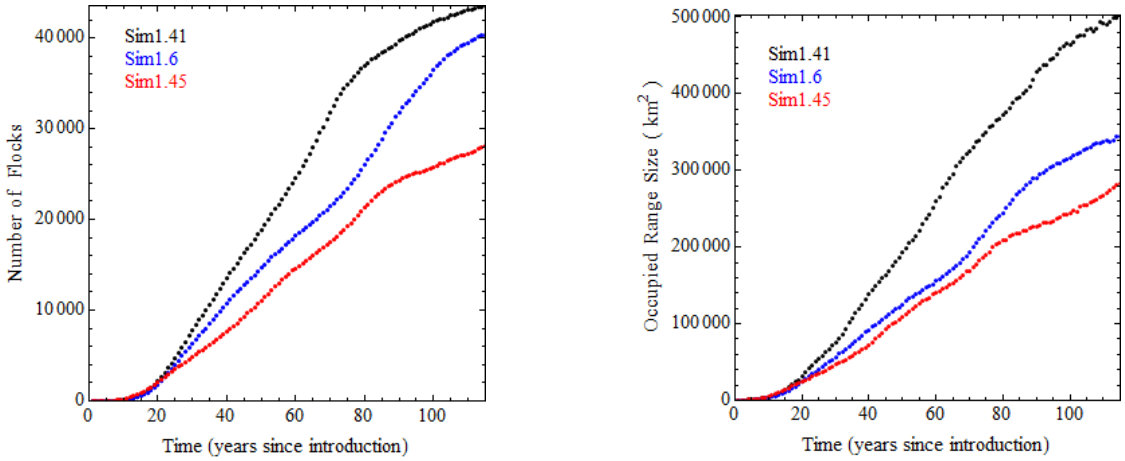


Figure 4.20: Predicted population growth and range size expansion of the European Starling for the three top performing models according to a combination of historical and current records' evaluations.

4.3.4 Parameter Sensitivity Analysis

The results shown so far of the top performing models gave the impression that models that incorporated habitat suitability map SDM1 performed generally better than those that incorporated SDM2. According to overall accuracy and kappa evaluation criteria, SDM1 performed significantly better than SDM2 (F-ratio = 44.097, $p \leq 0.05$ and F-ratio = 24.611, $p \leq 0.05$ respectively, Figure 4.21). However, according to the RMA slope and AUC evaluation criteria, SDM2 significantly outperformed SDM1 (F-ratio = 14.476, $p \leq 0.05$ and F-ratio = 8.897, $p \leq 0.05$ respectively) (Figure 4.21). It is not surprising that SDM2 performed better in AUC results because as we've seen, AUC is biased towards sensitivity and due to SDM2's greater proportion of high suitable habitat regions (Figure 4.22) the models were able to occupy a greater proportion of land mass than those employing SDM1 as dispersal is habitat dependent (density plots after 100 time steps for all 108 models are shown in Appendix G). The same holds true for better predicted slope results, as the greater proportion of high suitable regions fuels a faster range expansion leading to a RMA slope closer to one. Due to the better slope results, eight out of the top ten performing models in historical records' evaluations employed SDM2 (R^2 results were very similar between SDM1 and SDM2 models, A H). SDM1 performed better in current records evaluations as nine out of the top ten performing models employed SDM1, where the smaller proportion of high suitability areas encouraged slower range expansion and was therefore less likely to over predict the SABAP 1 distribution.

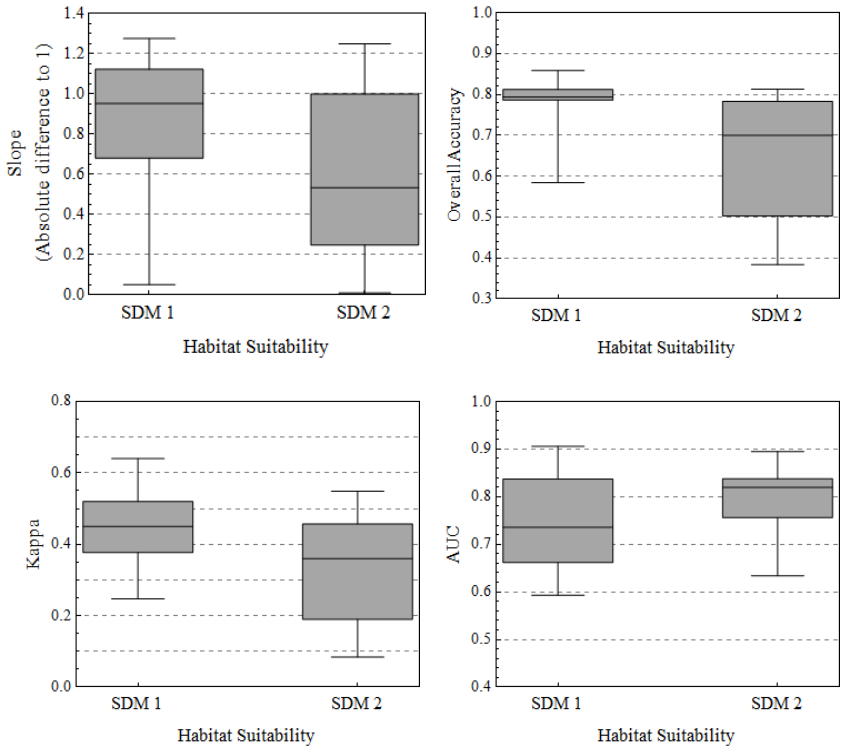


Figure 4.21: Box plots for varying the habitat suitability parameter among the 108 models and when evaluated according to those evaluation criteria showing significant differences. All box plots and parameter sensitivity analyses results are shown in A H.

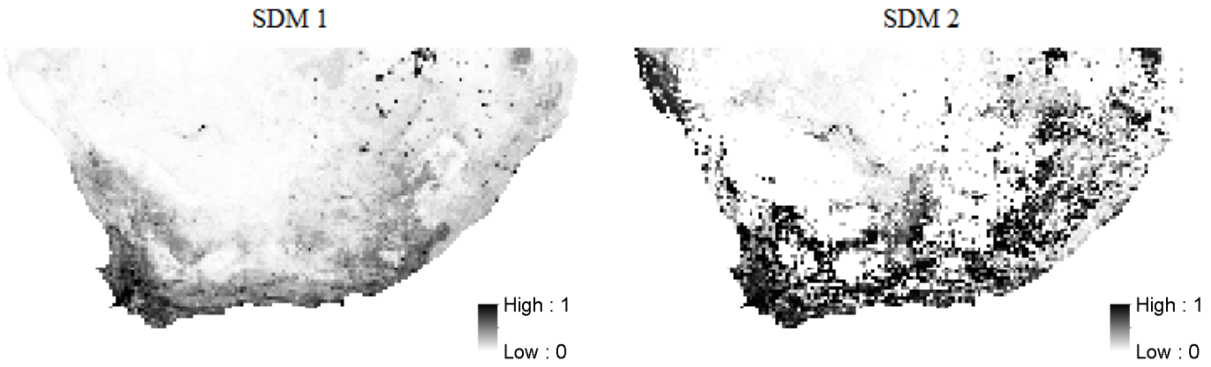


Figure 4.22: The two different habitat suitability maps employed in model simulations.

Different habitat suitability maps evidently have a great influence on model performance. SDM1 relied on *a priori* knowledge of important predictor variables which resulted in models that performed well when evaluating with current records. I then had a stepwise procedure of eliminating variables that gave rise to SDM2 which performed better when evaluating with historical records. Both methods of choosing predictor variables were important, but to different phases of the range expansion process as measured by different independent datasets. Winter precipitation contributed by far the most (73.2%) to SDM1's distribution while human footprint (44.9%) and annual mean temperature (30.4%) contributed the most

to SDM2's distribution. During the initial exploration phases of the invasion process it is plausible that human influences and temperature were leading decision makers in the distribution of the starlings whereas later on, during the establishment and further expansion phases, winter precipitation played a dominant role in predicting where the starlings were more likely to settle. The starlings, having a stronger tendency towards reproductive success during the early phase, would rely more on human influences for survival in terms of access to food and nesting sites. Additionally, during this early phase they reached many mountain barriers and in order to cross them they would most likely also depend on human influences in terms of road networks and established developments along the roads.

When combining both evaluation methods SDM1 performed best overall as eight out of the top ten models employed SDM1. The better performance shown by SDM1 in general suggests that winter precipitation played a leading role in predicting the starling's range expansion across Southern Africa. Applying *a priori* knowledge for choosing environmental predictors is therefore deemed highly successful.

The most noteworthy impacts that changing elevation access from 100 m to 300 m had were in significantly increasing RMA slope predictions (F-ratio = 15.783, $p \leq 0.05$) and significantly increasing AUC predictions (F-ratio = 5.707 $p \leq 0.05$) (Figure 4.23). Lifting geographic barriers were therefore important for allowing the starlings to disperse successfully and reach the desired distributions. The increase in model performance was more evident in a change of elevation access from 100 m to 200 m than from 200 m to 300 m (A H), suggesting that increasing elevation access does not lead to a linear increase in model performance, but that there exists a possible upper limit to elevation access restricting vertical flight dispersal.

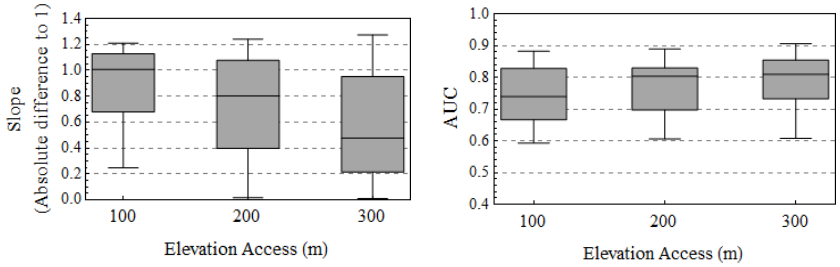


Figure 4.23: Box plots for varying the elevation access parameter among the 108 models and when evaluated according to those evaluation criteria showing significant differences. All box plots and parameter sensitivity analyses results are shown in A H.

Increasing the number of dispersal choices, specifically from five to ten, resulted in a significant decrease in model performance for slope (F-ratio = 8.527 $p \leq 0.05$), R^2 (F-ratio = 33.239, $p \leq 0.05$), TSS (F-ratio = 9.33, $p \leq 0.05$) and AUC (F-ratio = 23.007, $p \leq 0.05$) (Figure 4.24). Models employing ten dispersal choices frequently appeared among the three worst performing models across all evaluation criteria as this greatly restricted the starlings' movement and therefore their ability to match the

distribution of SABAP 1. Decisions based on simple comparisons between habitats are therefore more realistic for the starling's dispersal decision making procedure.

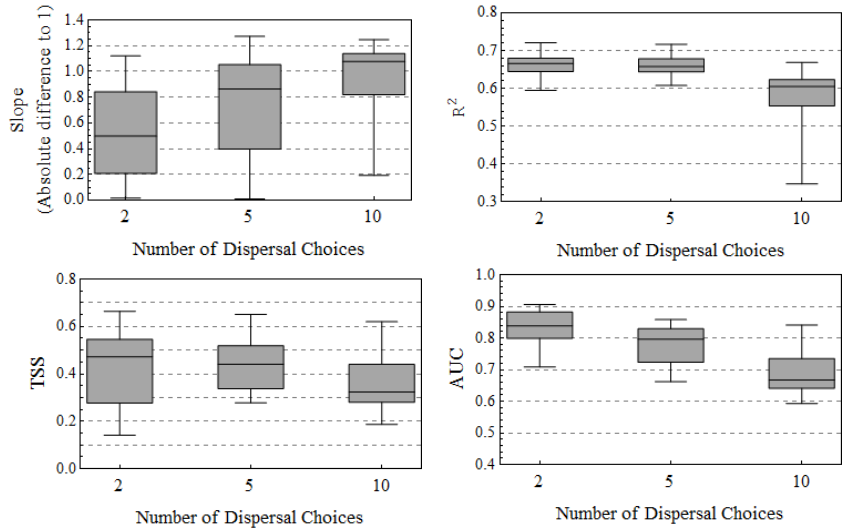


Figure 4.24: Box plots for varying the number of dispersal choices parameter among the 108 models and when evaluated according to those evaluation criteria showing significant differences. All box plots and parameter sensitivity analyses results are shown in A H.

Changing the dispersal function is also linked to the cognitive ability of the starlings and hence gave evaluation results similar to that of changing the number of dispersal choices. Significant decreases in model performance was evident for slope (F-ratio = 18.321, $p \leq 0.05$), TSS (F-ratio = 4.874, $p \leq 0.05$) and AUC results (F-ratio = 11.152, $p \leq 0.05$) when the dispersal decision was not only based on the habitat suitability, but also the distance from current to new locations (Figure 4.25). Once again simple comparisons between habitats dominated the starling's decision making capacity. Another explanation for why dispersal decisions based only on HS performed better is that the dispersal kernel has already considered distance, therefore making the additional consideration of distance in the dispersal function redundant.

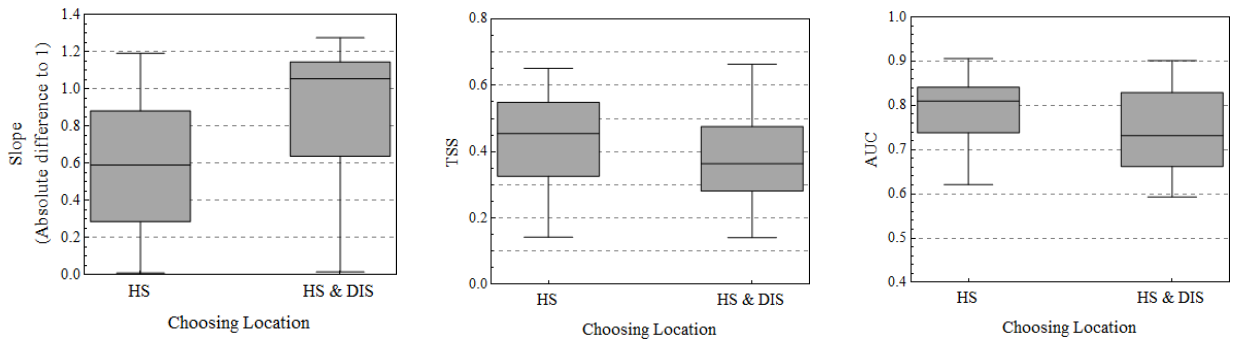


Figure 4.25: Box plots for varying the dispersal function parameter among the 108 models and when evaluated according to those evaluation criteria showing significant differences. All box plots and parameter sensitivity analyses results are shown in A H.

When changing the maximum dispersal distance from 100 km to 300 km, slope (F-ratio = 21.106, $p \leq 0.05$) and AUC (F-ratio = 10.434, $p \leq 0.05$) significantly increased while R^2 (F-ratio = 8.145, $p \leq 0.05$) and overall accuracy (F-ratio = 14.346, $p \leq 0.05$) significantly decreased (Figure 4.26). Unpredictable variation among results was also seen when studying changes with SDM1 and SDM2 separately (A H), with some changes leading to significant increases of evaluation results and others leading to significant decreases. Due to this instability in the results, it is uncertain what the change of dispersal distance on model performance is on its own, but it is more likely to play a significant role when combined with other parameters.

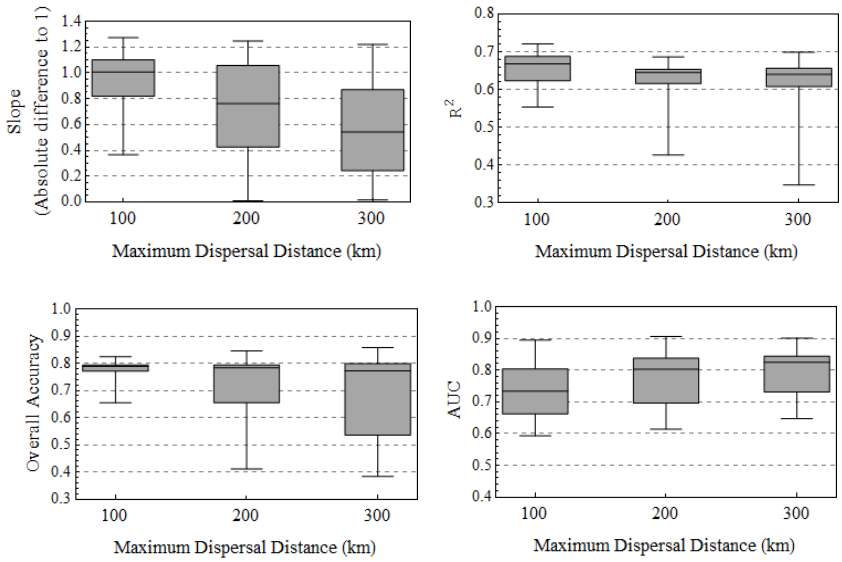


Figure 4.26: Box plots for varying the dispersal distance parameter among the 108 models and when evaluated according to those evaluation criteria showing significant differences. All box plots and parameter sensitivity analyses results are shown in A H.

4.3.5 Null Models

4.3.5.1 Null Model 1

Null model 1 excluded geographic barriers while keeping other parameters at their average settings (5 dispersal choices, maximum dispersal distance of 200 km, dispersal decision based only on HS). This null model was compiled for both habitat suitability maps and correspondingly compared with other models with the same parameter settings that only varied elevation access limits. In the graphical representations of the starling's relative density after 100 time steps (Figure 4.27 and Figure 4.28) we observe the influence the elevation barrier has on the starling's distribution. Removing elevation barriers by giving the flocks greater elevation access placed fewer restrictions on their movements which allowed for greater range expansions and distributions.

Null1.1 that incorporated SDM1 performed second worst (among the four comparative models involving SDM1, Table 4.7) with significantly different model performance according to R^2 (F-ratio = 135.061, $p \leq 0.05$, Table 4.8) and overall accuracy evaluations (F-ratio = 62.517, $p \leq 0.05$). Among those four models, Sim1.45, with 300 m elevation access, performed best. Null1.2 that incorporated SDM2 performed worst (among the four comparative models involving SDM2, Table 4.7) while Sim2.9, with 100 m elevation access, performed best. Comparing Figure 4.27 and Figure 4.28 it is evident the enormous impact that different habitat suitability maps have on the starling's distribution. It is visually apparent among all the distributions why Sim1.45 and Sim2.9 lead to the best model performances. The reason why both a model with a 100 m and one with a 300 m elevation access could perform so well is due to the different habitat suitability maps. When dispersing, a flock based their decision of where to go on the most suitable areas; therefore the habitat suitability map employed will largely impact the ultimate distribution. Furthermore, mortality rates were also dependent on habitat suitability. For these reasons SDM1 greatly limited the starling's dispersal ability due to the much lower proportion of high suitable areas than that of SDM2 (Figure 4.22). Incorporating an elevation barrier in the model is necessary to define the dispersal routes of the species; however it should be done with care depending on the habitat suitability map used.

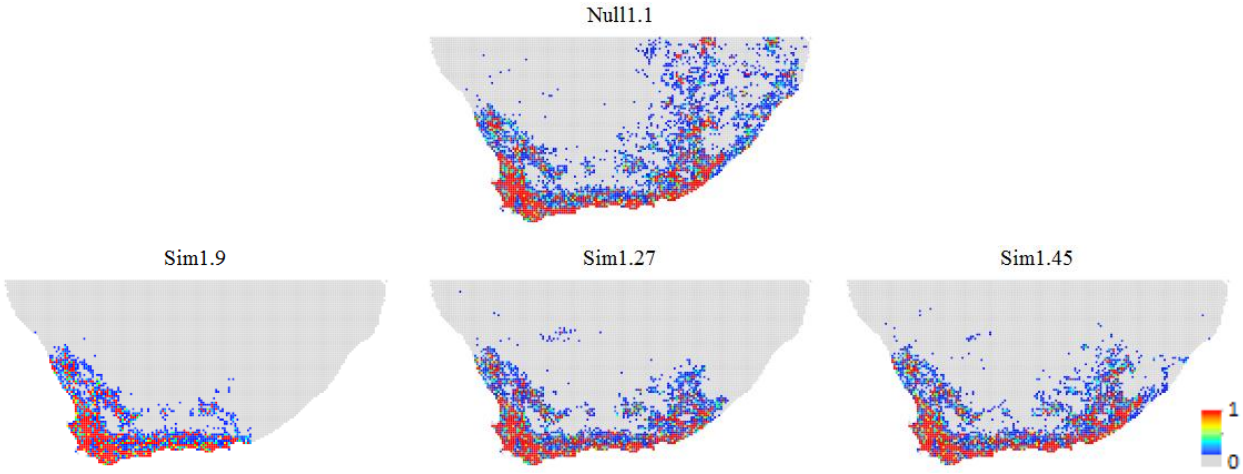


Figure 4.27: The European Starling's relative density after 100 simulated time steps for null model 1 (unlimited elevation access) with habitat suitability map SDM1 (Null1.1) compared with models varying in elevation access, but otherwise using the same parameter settings (CH5, 5 to 200 km, HS). From left to right: Sim1.9 has 100 m elevation access, Sim1.27 has 200 m and Sim1.45 has 300 m.

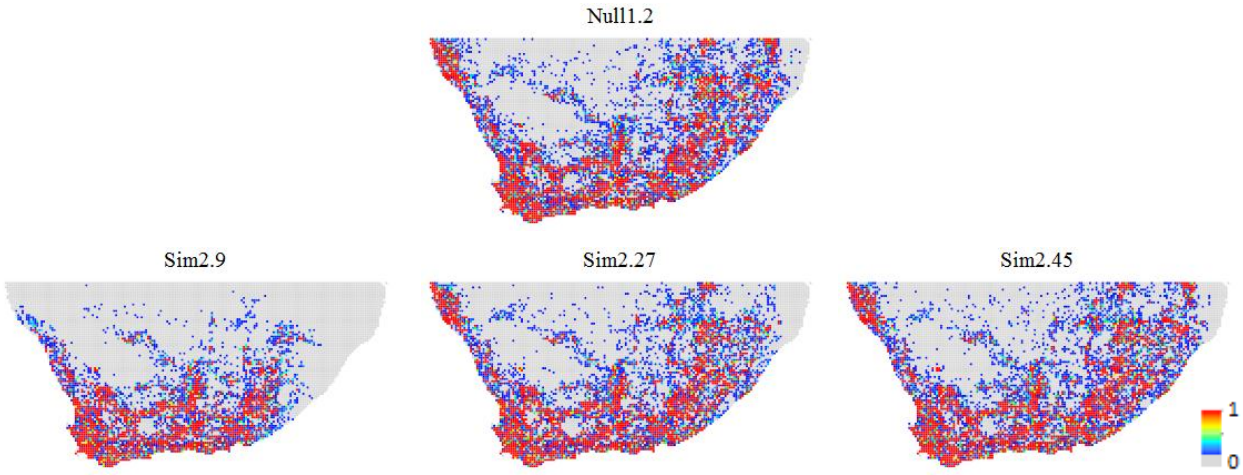


Figure 4.28: The European Starling's relative density after 100 simulated time steps for null model 1 (unlimited elevation access) with habitat suitability map SDM2 (Null1.2) compared with models varying in elevation access, but otherwise using the same parameter settings (CH5, 5 to 200 km, HS). From left to right: Sim2.9 has 100 m elevation access, Sim2.27 has 200 m and Sim2.45 has 300 m.

4.3.5.2 Null Model 2

Null model 2 excluded the two parameters related to cognitive ability: dispersal choices and therefore also the dispersal function. The remaining parameters were kept at their average settings (200 m elevation access, maximum dispersal distance of 200 km) and were compared to the models with corresponding parameter settings which varied only in their degree of cognitive ability.

Increasing cognitive ability, through increasing the number of dispersal choices and taking into consideration both HS and distance when dispersing, lead to a more careful and restricted dispersal movement and therefore a slower range expansion (Sim1.27, Sim1.28, Sim1.33 and Sim1.34 in Figure 4.29 and Sim2.27, Sim2.28, Sim2.33 and Sim2.34 in Figure 4.30). Instead of just random dispersal, the flocks dispersed to the most suitable locations when more options were considered with increased cognitive ability.

For habitat suitability map SDM1 the null model performed best among all ten comparative models (Table 4.8). The starlings had limited restriction on their movement in this null model, dispersing randomly and were only restricted by a 200 m elevation barrier, a maximum dispersal of 200 km and habitat dependent mortality. The starlings were unable to reach the observed distributions when adding further dispersal restrictions through more complex cognitive abilities. The limited high suitability areas of SDM1 further restricted their movement. The fact that Null2.1, which incorporated SDM1, performed so well could imply that SDM1 in itself is a good indication of suitable habitat. Even without cognitive ability their habitat dependent mortality alone would accurately regulate their distribution.

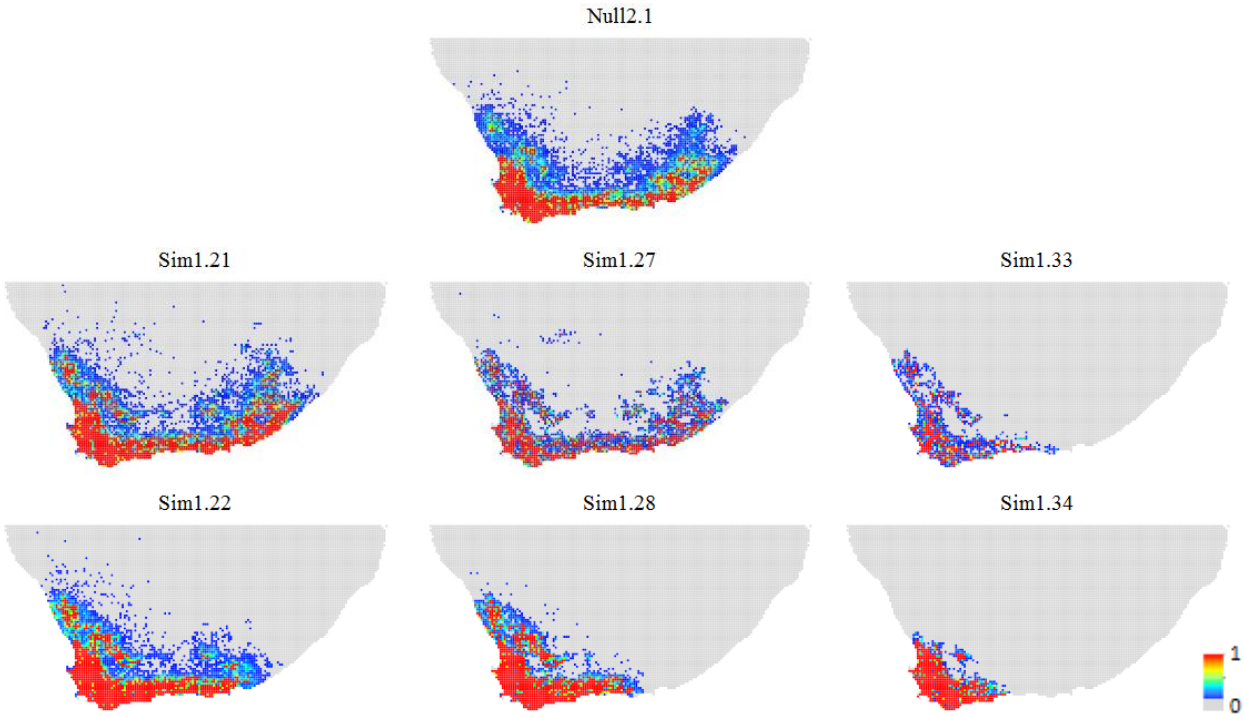


Figure 4.29: The European Starling's relative density after 100 simulated time steps for null model 2 (no cognitive ability related parameters) with habitat suitability map SDM1 (Null2.1), compared with models varying in cognitive ability levels, but otherwise using the same parameter settings (DEM200, 5 to 200 km). Sim1.21 and Sim1.22 had two dispersal choices, Sim1.27 and Sim1.28 had five dispersal choices while Sim1.33 and Sim1.34 had ten dispersal choices. Sim1.21, Sim1.27 and Sim1.33 dispersed according to HS alone while Sim1.22, Sim1.28 and Sim1.34 dispersed according to HS and distance.

Null2.2 on the other hand, with SDM2 as a habitat suitability map, was ranked fourth among the ten comparative models. Sim2.28 was ranked best; it had five dispersal choices and dispersed according to both HS and distance. The extremely high suitable areas for SDM2 allowed the starling to once again survive and establish across a greater range and therefore more restrictions were necessary to obtain the desired distributions.

The habitat dependent mortality rate limited the starling's establishment ability in SDM1 models more than in SDM2 models due to differing degrees of suitable habitat. In SDM1 models the flocks established in locations with an average suitability which implied that not all flocks would survive and in the subsequent time step there were fewer starlings to disperse, thereby hindering a rapid expansion. SDM2 models on the other hand allowed the flocks to reach greater suitable habitat sites which allowed the majority of flocks to survive and reproduce, therefore leading to a greater expansion rate. As a result, I have two possible scenarios: one where SDM1 alone without cognitive ability lead to good model performance, compared to one where SDM2 required a certain degree of restriction through increasing cognitive ability to lead to good model performance. Depending on the HS map, more or less restrictions were required to increase model performance.

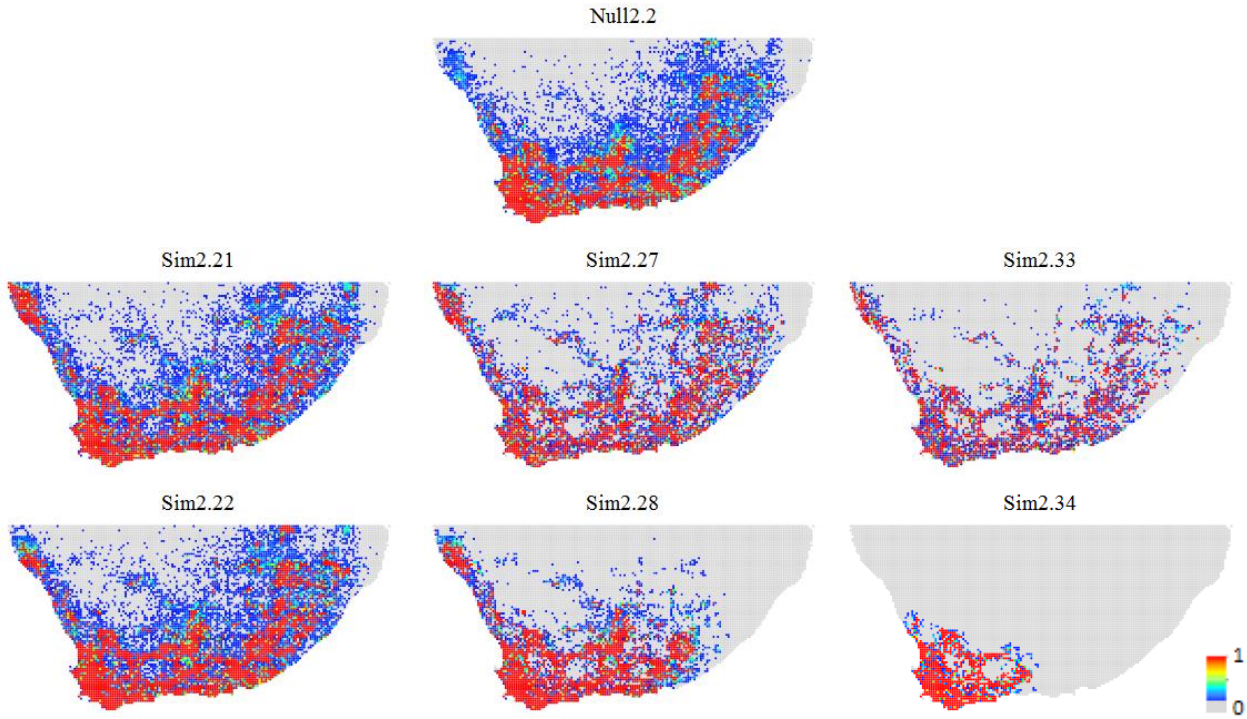


Figure 4.30: The European Starling's relative density after 100 simulated time steps for null model 2 (no cognitive ability related parameters) with habitat suitability map SDM2 (Null2.2), compared with models varying in cognitive ability levels, but otherwise using the same parameter settings (DEM200, 5 to 200 km). Sim2.21 and Sim2.22 had two dispersal choices, Sim2.27 and Sim2.28 had five dispersal choices while Sim2.33 and Sim2.34 had ten dispersal choices. Sim2.21, Sim2.27 and Sim2.33 dispersed according to HS alone while Sim2.22, Sim2.28 and Sim2.34 dispersed according to HS and distance.

4.3.5.3 Null Model 3

The final null model had neither elevation barriers nor cognitive ability related parameters to discern between locations. Dispersal restrictions were based solely on a maximum dispersal distance of 200 km and habitat dependent mortality. This null model was therefore compared to those models that had the same habitat suitability map and a maximum dispersal distance of 200 km. Null3.1 which incorporated SDM1 ranked 10th out of the 19 comparative models, while Sim1.45 performed best (the same model that performed best across all evaluations among all 108 models). Sim1.45 had a 300 m elevation access and five dispersal choices which was dependent on HS, restricting the starling's distribution just enough to reach the observed distributions (Figure 4.31).

Null3.2 which incorporated SDM2 was ranked as the worst performing model out of the 19 comparative models. Sim2.28 performed best with elevation access of 200 m and five dispersal choices dependent on both HS and distance. These parameter settings restricted the starling's distribution to a greater degree than when using Sim1.45, as the habitat suitability map SDM2 required greater restrictions to reach the desired distributions (Figure 4.32). Once again it can be concluded that more or less restrictions were required to increase model performance, depending on the HS map.

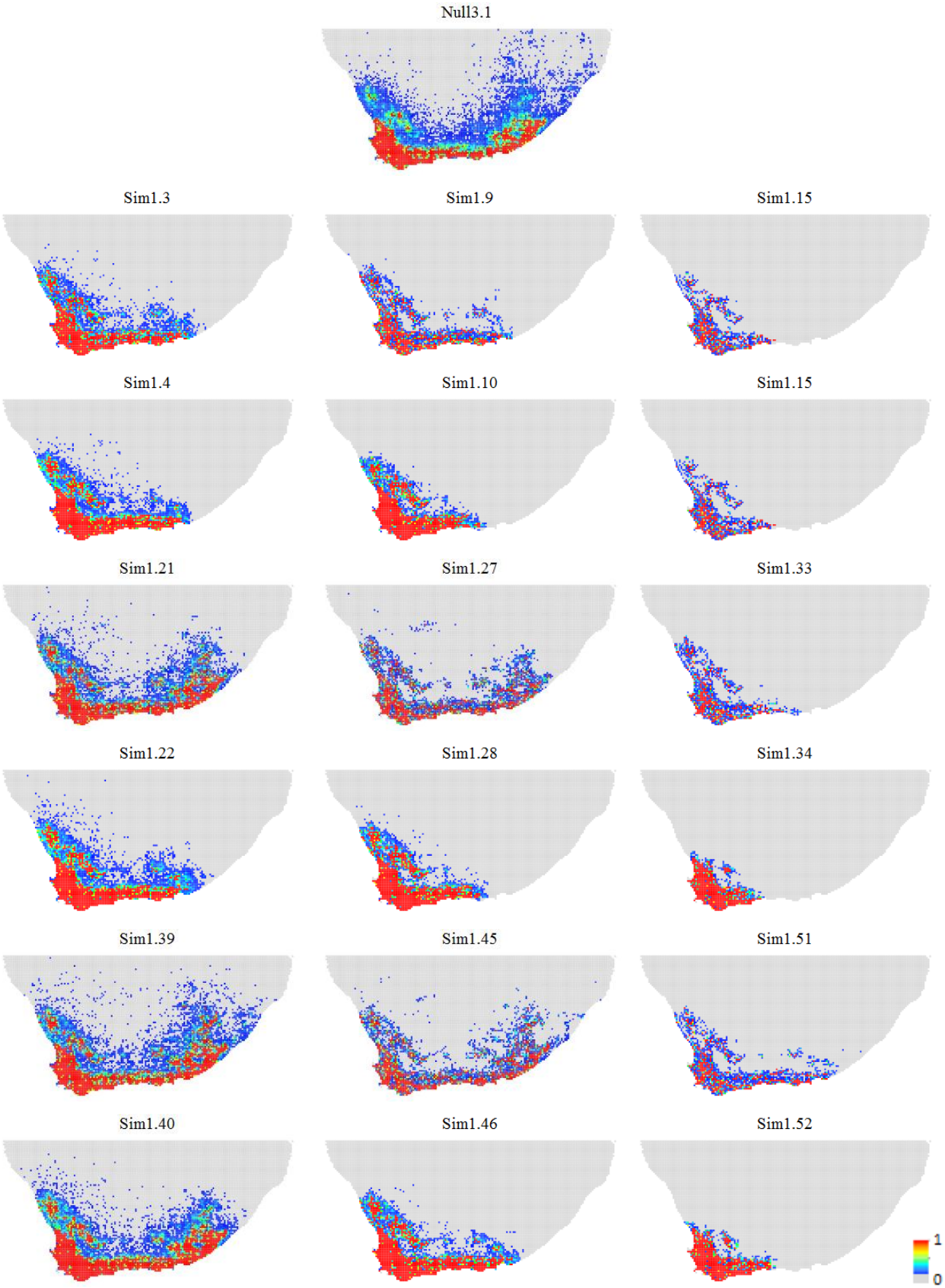


Figure 4.31: The European Starling's relative density after 100 simulated time steps for null model 3 (without any elevation barriers or cognitive ability related parameters) with habitat suitability map SDM1 (Null3.1), compared with models varying in cognitive ability levels and elevation access, but with the same maximum dispersal distance of 200 km.

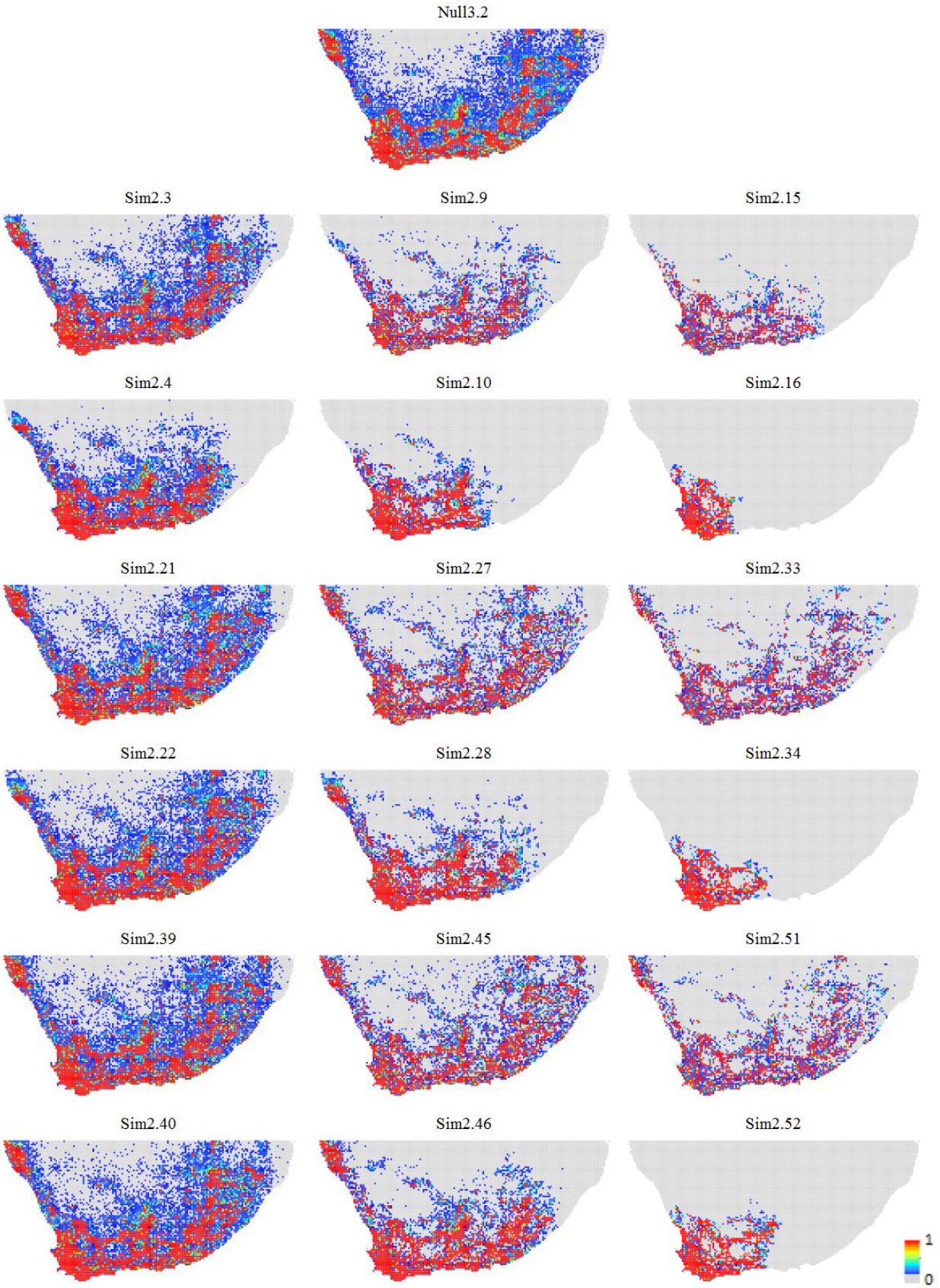


Figure 4.32: The European Starling's relative density after 100 simulated time steps for null model 3 (without any elevation barriers or cognitive ability related parameters) with habitat suitability map SDM2 (Null3.2), compared with models varying in cognitive ability levels and elevation access, but with the same maximum dispersal distance of 200 km.

Table 4.7: Evaluation results and rankings for each null model and its comparative group of models.

	Historical Records' Evaluations			Current Records' Evaluations				Rankings		
Model	Slope	Absolute Difference of Slope to 1	R^2	Overall Accuracy	Kappa	TSS	AUC	Historical Records' Evaluations	Current Records' Evaluations	All Evaluations (Slope, R^2, Kappa & TSS)
Null1.1	0.993	0.007	0.424	0.679	0.368	0.466	0.829	2	3.5	3
Sim1.9	1.871	0.871	0.651	0.813	0.492	0.452	0.736	3.5	3.5	4
Sim1.27	1.554	0.554	0.644	0.846	0.621	0.629	0.837	3.5	1.5	2
Sim1.45	1.428	0.428	0.677	0.830	0.602	0.639	0.854	1	1.5	1
Null1.2	0.578	0.422	0.652	0.486	0.172	0.267	0.803	2	4	4
Sim2.9	1.625	0.625	0.639	0.709	0.422	0.531	0.853	4	1	1
Sim2.27	1.160	0.160	0.63	0.542	0.230	0.341	0.820	3	2	3
Sim2.45	0.991	0.009	0.639	0.503	0.190	0.290	0.806	1	3	2
Null2.1	1.541	0.541	0.662	0.817	0.597	0.678	0.895	1	1	1
Sim1.21	1.465	0.465	0.624	0.784	0.540	0.632	0.889	2	3	2.5
Sim1.22	1.843	0.843	0.648	0.789	0.500	0.527	0.799	4	4	4
Sim1.27	1.554	0.554	0.644	0.846	0.621	0.629	0.837	4	2	2.5
Sim1.28	2.180	1.180	0.65	0.786	0.410	0.372	0.697	6	5.5	5
Sim1.33	2.077	1.077	0.582	0.795	0.397	0.335	0.671	7	5.5	7
Sim1.34	2.107	1.107	0.654	0.791	0.340	0.268	0.635	4	7	6
Null2.2	1.155	0.155	0.637	0.488	0.181	0.282	0.869	2	4.5	4
Sim2.21	1.032	0.032	0.653	0.427	0.123	0.201	0.829	1	7	4
Sim2.22	1.192	0.192	0.632	0.437	0.133	0.216	0.847	4.5	6	7
Sim2.27	1.160	0.160	0.630	0.542	0.23	0.341	0.820	4.5	4.5	4
Sim2.28	1.786	0.786	0.647	0.718	0.424	0.519	0.829	4.5	1	1
Sim2.33	1.456	0.456	0.500	0.669	0.357	0.46	0.806	7	2.5	6
Sim2.34	2.112	1.112	0.650	0.794	0.399	0.340	0.674	4.5	2.5	2
Null3.1	0.906	0.094	0.584	0.694	0.418	0.55	0.913	8.5	11	10
Sim1.3	1.901	0.901	0.678	0.804	0.511	0.512	0.780	5	5	7
Sim1.4	1.918	0.918	0.667	0.789	0.471	0.469	0.760	6	12	9
Sim1.9	1.871	0.871	0.651	0.813	0.492	0.452	0.736	11	9.5	12
Sim1.10	2.129	1.129	0.658	0.794	0.422	0.375	0.696	14.5	13	13
Sim1.15	2.190	1.190	0.604	0.792	0.367	0.301	0.653	19	17	18
Sim1.16	2.162	1.162	0.607	0.783	0.297	0.228	0.614	17.5	19	19
Sim1.21	1.465	0.465	0.624	0.784	0.540	0.632	0.889	8.5	3	5
Sim1.22	1.843	0.843	0.648	0.789	0.500	0.527	0.799	12	8	8
Sim1.27	1.554	0.554	0.644	0.846	0.621	0.629	0.837	8.5	2	4
Sim1.28	2.180	1.180	0.650	0.786	0.410	0.372	0.697	16	15.5	15
Sim1.33	2.077	1.077	0.582	0.795	0.397	0.335	0.671	17.5	14	17
Sim1.34	2.107	1.107	0.654	0.791	0.340	0.268	0.635	14.5	18	16
Sim1.39	1.125	0.125	0.671	0.726	0.461	0.585	0.906	1	9.5	3
Sim1.40	1.404	0.404	0.651	0.778	0.533	0.632	0.895	4	4	2
Sim1.45	1.428	0.428	0.677	0.830	0.602	0.639	0.854	2	1	1
Sim1.46	1.868	0.868	0.679	0.812	0.511	0.491	0.762	3	6	6

Sim1.51	1.738	0.738	0.616	0.824	0.511	0.458	0.735	13	7	11
Sim1.52	2.058	1.058	0.660	0.801	0.385	0.311	0.657	8.5	15.5	14
Null3.2	0.782	0.218	0.606	0.427	0.125	0.205	0.848	16	16	19
Sim2.3	1.471	0.471	0.644	0.492	0.179	0.277	0.838	9	13	16
Sim2.4	1.552	0.552	0.642	0.635	0.331	0.451	0.868	13	6	10
Sim2.9	1.625	0.625	0.639	0.709	0.422	0.531	0.853	15	1	4
Sim2.10	2.054	1.054	0.653	0.772	0.449	0.463	0.757	8	5	3
Sim2.15	1.876	0.876	0.579	0.813	0.537	0.543	0.786	19	2	6
Sim2.16	2.140	1.140	0.613	0.783	0.334	0.272	0.638	18	12	18
Sim2.21	1.032	0.032	0.653	0.427	0.123	0.201	0.829	3	17	14
Sim2.22	1.192	0.192	0.632	0.437	0.133	0.216	0.847	10.5	14	17
Sim2.27	1.160	0.160	0.630	0.542	0.230	0.341	0.820	10.5	11	12.5
Sim2.28	1.786	0.786	0.647	0.718	0.424	0.519	0.829	7	3	1
Sim2.33	1.456	0.456	0.500	0.669	0.357	0.460	0.806	17	10	11
Sim2.34	2.112	1.112	0.650	0.794	0.399	0.340	0.674	5	8	5
Sim2.39	0.982	0.018	0.683	0.411	0.109	0.180	0.828	2	19	15
Sim2.40	1.082	0.082	0.687	0.418	0.115	0.188	0.839	1	18	12.5
Sim2.45	0.991	0.009	0.639	0.503	0.190	0.290	0.806	6	15	9
Sim2.46	1.524	0.524	0.642	0.677	0.391	0.517	0.857	4	4	2
Sim2.51	1.426	0.426	0.427	0.655	0.350	0.462	0.813	14	9	8
Sim2.52	2.248	1.248	0.548	0.784	0.399	0.359	0.688	12	7	7

Table 4.8: ANOVA significance tests of model performance among each null model and its comparative group of models according to each evaluation criteria. Significant differences in sample means are highlighted in red. The relatively few significant differences are due to small sample sizes which lead to large critical values for the F-test (F_{crit}). F_{crit} refers to the relevant critical value used for the F-test, F refers to the F-ratio and P refers to the p-value.

		RMA Slope		R^2		Overall Accuracy		Kappa		TSS		AUC	
	F_{crit}	F	P	F	P	F	P	F	P	F	P	F	P
Null 1.1	18.5	5.368	0.146	**135.061**	**0.007**	**62.517**	**0.016**	6.414	0.127	0.781	0.470	0.074	0.811
Null 1.2	18.5	0.180	0.713	7.111	0.117	0.610	0.517	0.576	0.527	0.673	0.498	0.701	0.491
Null 2.1	6.61	1.015	0.360	0.916	0.382	0.527	0.500	1.311	0.304	1.650	0.255	1.641	0.256
Null 2.2	6.61	0.444	0.535	0.083	0.785	0.445	0.534	0.447	0.533	0.218	0.660	0.986	0.366
Null 3.2	4.45	**5.147**	**0.037**	**4.569**	**0.047**	**15.594**	**0.001**	0.272	0.609	0.451	0.511	2.735	0.117
Null 3.3	4.45	0.575	0.459	0.027	0.871	1.744	0.204	1.688	0.211	1.517	0.235	0.518	0.481

4.3.6 Forecasting Future Distributions

The three top performing models according to historical records' evaluations (Sim1.41), current records' evaluations (Sim1.6) and both historical and current records' evaluations combined (Sim1.45) were used to forecast potential future range expansions of the starling across Southern Africa (Figure 4.33, Figure 4.34 and Figure 4.35).

Sim1.41 (SDM1, DEM300, CH2, 5 to 300 km, HS) had limited dispersal restrictions and thus a fast range expansion in order to perform well when compared to historical records' evaluations. According to the model's future prediction, most of South Africa is already invaded by the starlings by 2030 and they've reached a saturation phase after which the only dispersal events are towards neighbouring Zimbabwe and Mozambique (Figure 4.33). Highest densities occur along the Southern coastlines due to high suitability areas as predicted by winter precipitation. However, due to the irregularities in the historical records' data collection (as discussed in Section 4.3.1), I would not recommend this forecast as one to base management planning on as this model and its evaluation performance should be treated with care. Additionally, predictions that were made up to the year 2012 had a smaller modelling extent than predictions made from 2013 onwards. In Sim1.41 the flocks had already reached this northern boundary in 1980 restricting further dispersal (Figure 4.33), another reason why this forecast was not recommended.

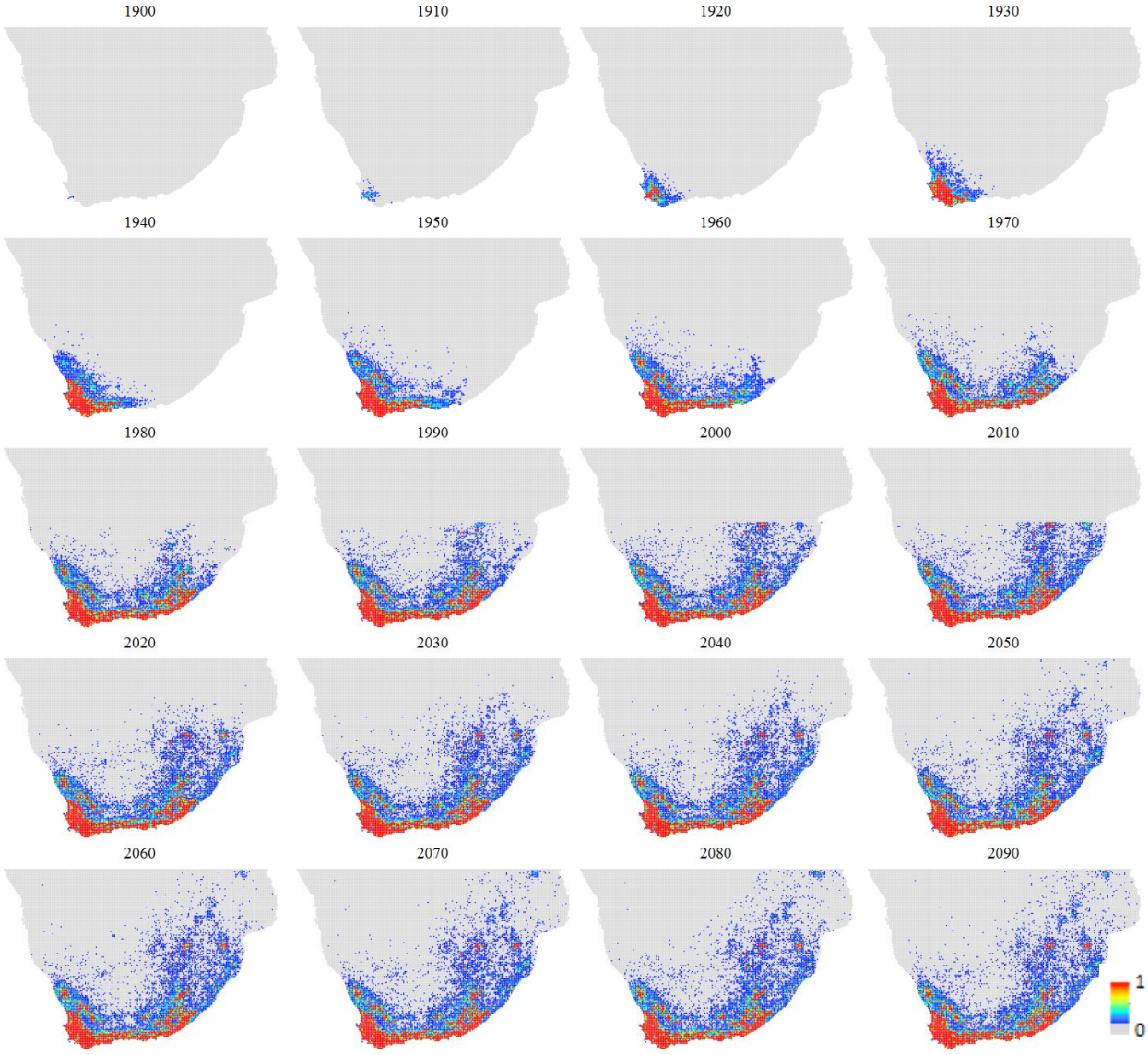


Figure 4.33: Past predictions and future projections of the European Starling's range expansion in Southern Africa from 1900 to 2090 according to the best ranked model in historical records' evaluations, Sim1.41, presented here with 10-year intervals.

Sim1.6 (SDM1, DEM100, CH2, 5 to 300 km, HS DIS) was more restricted than Sim1.41, both in terms of elevation access and the dispersal function. Consequently, its forecasted range expansion was also more restricted and the flocks first reached the Gauteng province around the year 2050. This is unrealistic as I know already from the SABAP 2 data that the European Starling has been sighted in Gauteng by the year 2012 (Figure 4.36). This model therefore runs the risk of potentially under predicting the true distribution of the starling which would be very misleading for future planning. Since I am dealing with an invasive species, under prediction is definitely not desired.

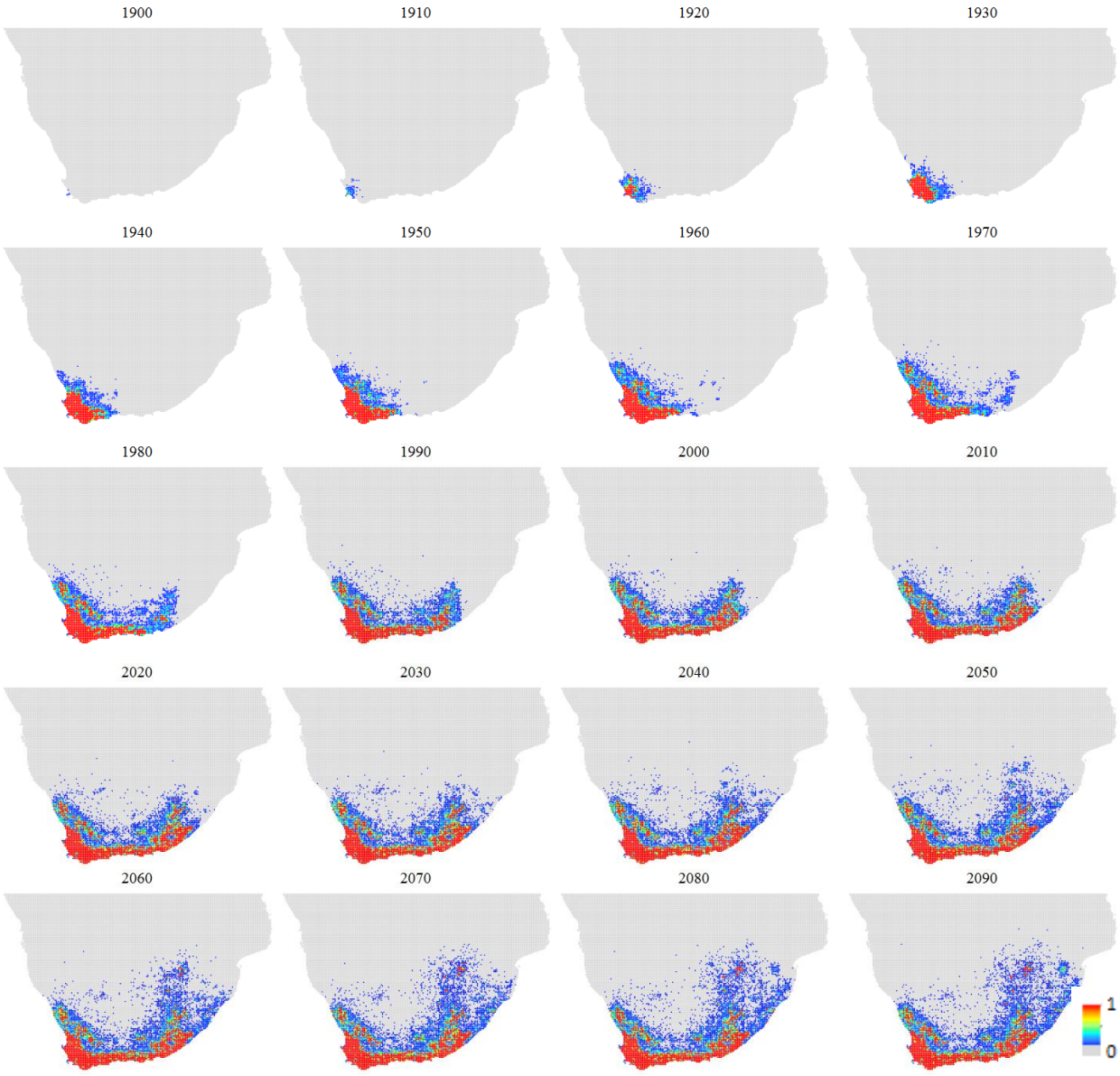


Figure 4.34: Past predictions and future projections of the European Starling's range expansion in Southern Africa from 1900 to 2090 according to the best ranked model in current records' evaluations, Sim1.6, presented here with 10-year intervals.

Sim1.45 (SDM1, DEM300, CH5, 5 to 200 km, HS), similar to Sim1.41 with an increased degree of cognitive ability, greatly restricted the starling's dispersal into the Karoo region, but allowed for ample

movement through the provinces of KwaZulu-Natal and Gauteng as well as into neighbouring Zimbabwe and Mozambique (Figure 4.35). Of all three models used for future projections, Sim1.45 most accurately projected the starling's presences in 2012 as observed in SABAP 2 (Figure 4.36). SABAP 2 was not used during evaluation methods because it has already been used when developing the habitat suitability maps in Chapter 3. I display it here as a visual comparison to my model predictions. SABAP 2 was the most recent data I had available regarding the starling's distribution giving us, at a finer resolution than SABAP 1, a good indication of where the starlings were present. Furthermore, Sim1.45 was chosen as top performing model according to an assortment of criteria and not restricted to evaluation with a single data set. The projections from this model as presented in Figure 4.35 were therefore considered most realistic out of all three models' projections and were subsequently defined as my optimal model.

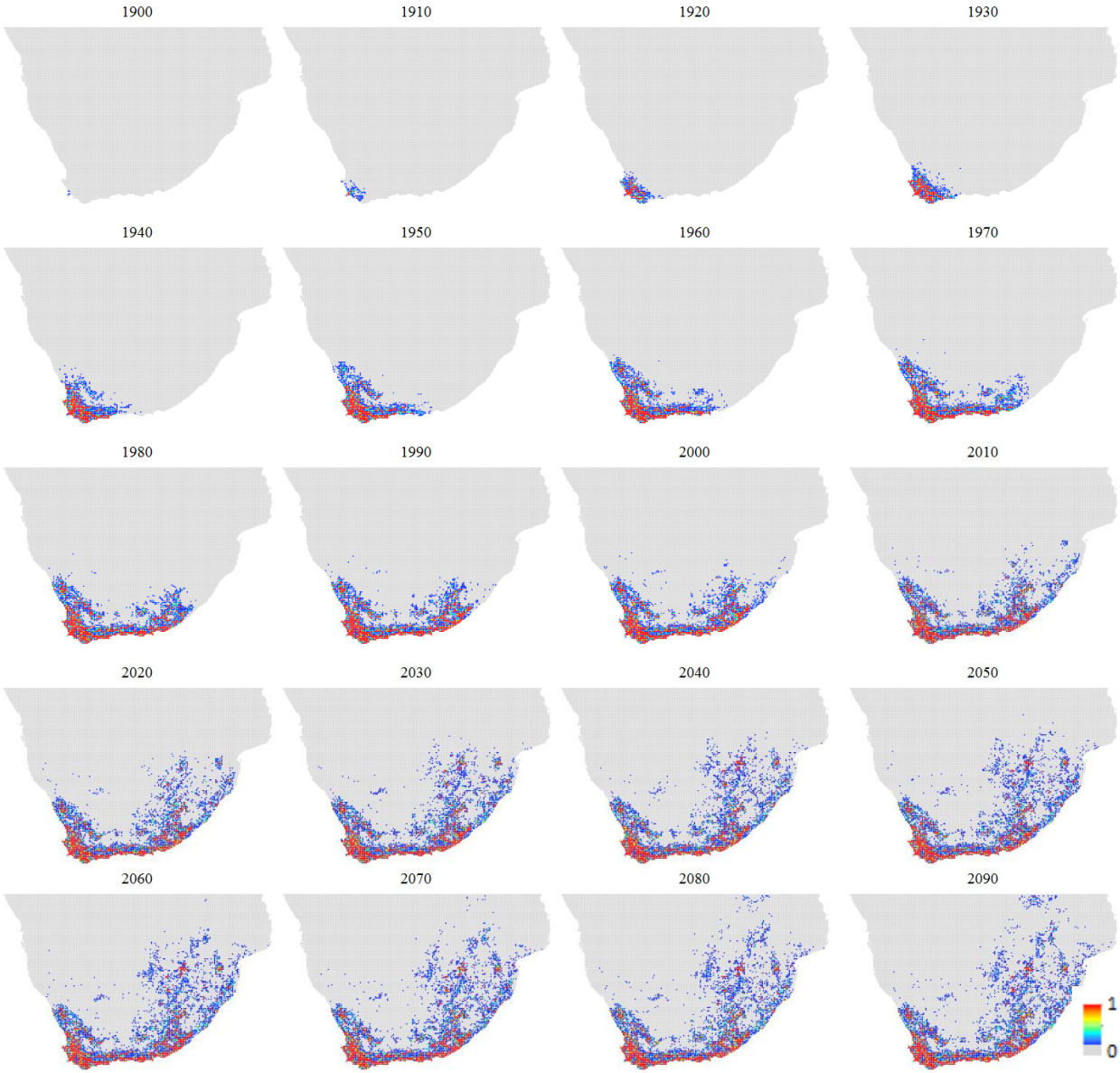


Figure 4.35: Past predictions and future projections of the European Starling's range expansion in Southern Africa from 1900 to 2090 according to the best ranked model in historical records and current records' evaluations, Sim1.45, presented here with 10-year intervals.

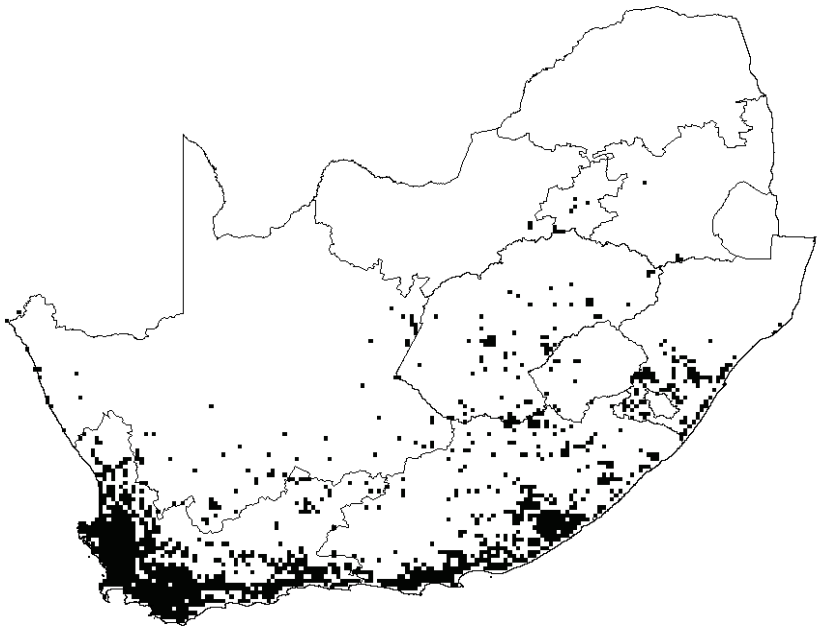


Figure 4.36: SABAP 2 distribution of European Starling presences in South Africa in 2012.

According to this future projection, limited dispersal will occur further along the western coastline through neighbouring Namibia or inland through the Great Karoo region, while further expansion is shown eastwards through the province of KwaZulu-Natal and north through the provinces of Free State, Gauteng, North West, Limpopo and Mpumalanga. Those are the areas that will need management planning and control measures if any further invasions by the starlings want to be restricted.

A detailed map showing this optimal model's prediction of the European Starling's dispersal route is shown in Figure 4.37. The background map displays the elevation where certain well known geographic features are noticeable and indicated on the map in blue circles (a – k). The starling's predicted routes from its introduction in 1897 to 2012 are indicated with green arrows and their projected routes from 2013 to 2100 are indicated with red arrows. The years specify when the model predicted the flocks' arrival at selected locations. Their initial dispersal route continued along the main N2 and N1 highways that went around the Hottentots Holand and Boland mountain ranges (a). In the north they went around the Cederberg and Skurweberg mountains (b). They took two different dispersal routes eastwards around the Langeberg mountains (c), one north and one south of the mountain ranges. They continued on these routes around the Tsitsikamma and Kouga mountains (e). They further on also avoided the Swartberge (d), Winterberge (h), Drakensberge (i), Maluti mountains in Lesotho (j) and the Waterberge in Limpopo province (k). The starlings were also able to avoid the semi-desert regions of the Little Karoo (f) and Great Karoo (g).

By the year 2100 there are certain regions where the starlings have settled in great numbers, forming sink populations. These regions include the entire coastline of South Africa as well as a number of large towns and cities of which the key ones are indicated on the map in orange circles (l - r).

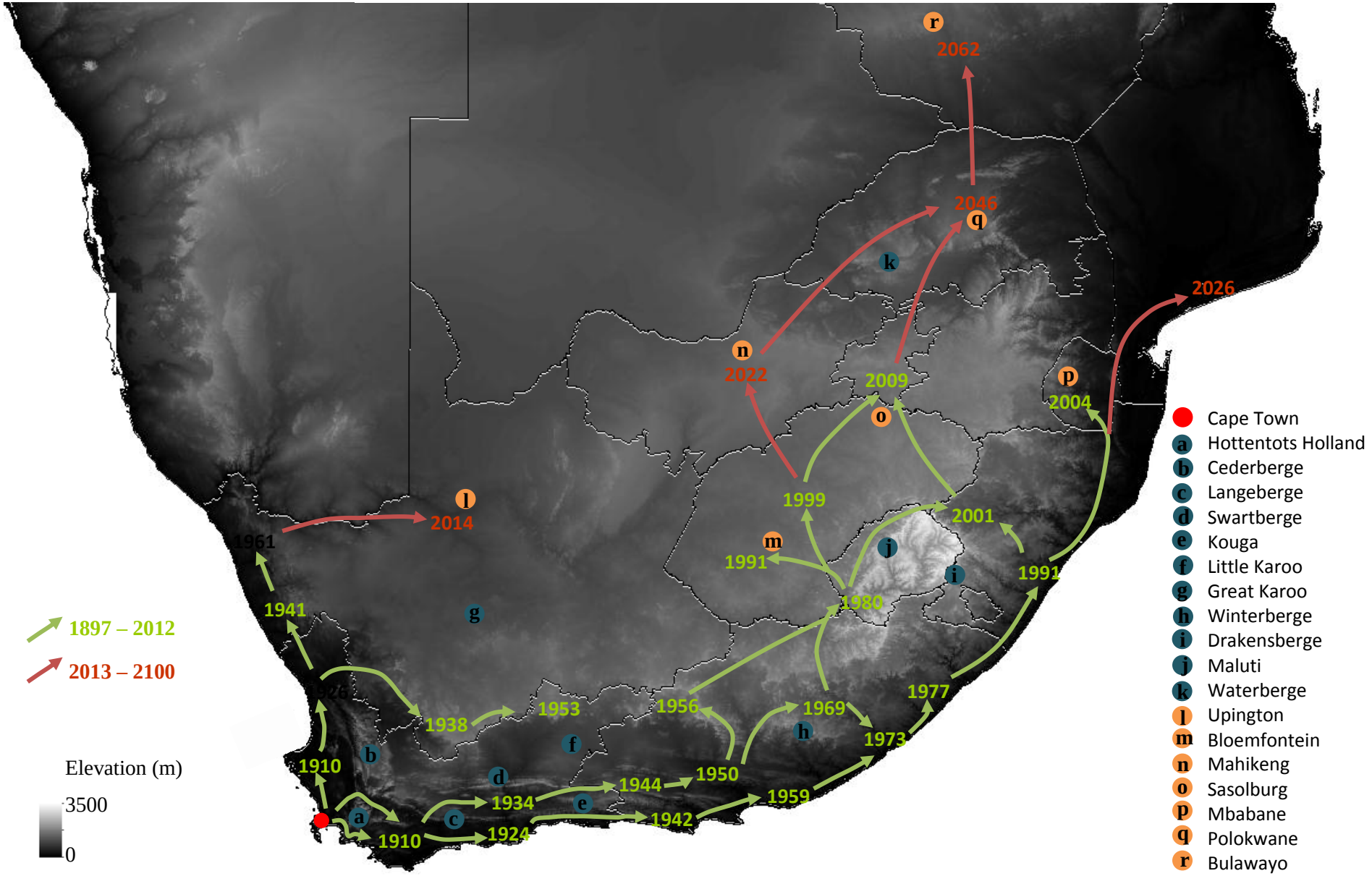


Figure 4.37: A schematic depiction of the European Starling's dispersal routes as predicted by the optimal model, Sim1.45.

4.3.7 Concluding Remarks

Designing a species' distribution model should be done with great care as it has an enormous impact on the dynamic model's outcomes. The better performance shown by SDM1 in general suggested that winter precipitation played a leading role in predicting the starling's distribution across Southern Africa. Applying *a priori* knowledge in choosing environmental predictors was therefore deemed highly successful. Nevertheless, we've only studied two different SDM techniques for defining HS maps, so further research is required in developing appropriate techniques for distribution modelling. A greater variety of HS maps should be studied to observe in greater detail and depth SDM impacts on the dynamic modelling outcomes.

Neither Sim1.6 nor Sim1.41, which were evaluated with historical records and current records respectively, gave as good projections as Sim1.45. This strengthens the concept that combining various evaluation techniques and data sets are desirable for finding an optimal model.

A problem that I encountered in the modelling procedure, especially with those scenarios where fast range expansion was evident, was that the northern boundary in my modelling framework acted as a reflecting boundary. As soon as the flocks reached it, they were reflected back south and could not cross which was highly unrealistic as land still persists across this boundary line. For the initial modelling procedure I was mainly interested in what happens along the Southern range expansions and not all simulations even reached the northern boundary within the 115 time steps. When evaluating with historical records this was not a problem as the towns I evaluated against were along the Southern coastlines only. When evaluating with current records, the whole extent was taken into consideration and a problem that occurred was an increase in relative density along the northern boundary due to reflection. Some simulations reached that boundary and had to be interpreted with care. For future predictions where I was interested in what happens around that boundary, I further expanded the northern boundary further north to avoid any such boundary problems.

A possible reason why my model was unable to predict the two-phase range expansion observed in the historical records is that spread rates are likely dominated by processes other than those incorporated in the model. Two potential mechanisms that were not implemented in the model are adaptation of dispersal traits and the spatial sorting of individuals with different dispersal abilities. My model therefore assumed homogenous dispersal which restricted the starling's ability to adapt their dispersal strategies during the invasion process.

Irrespective of the modelling method employed, it is often the case that when modelling species with fewer restricted ecological requirements, it is harder to portray their distributions accurately than when modelling species with greater restricted requirements (Brotons *et al*., 2004). The European Starling is one such resourceful species that can flourish in a variety of environments, making it more complicated to model accurately and for making future predictions.

CHAPTER 5

CONCLUSIONS AND FUTURE RECOMMENDATIONS

5.1 Conclusions

This study highlighted the importance of integrating various modelling techniques with biological knowledge to obtain an appropriate model for studying an invasive species' range expansion and the associated distribution dynamics. All my objectives were carried out to great satisfaction; a short summary of the achievements is as follow:

Objective I: I constructed a potential distribution for the European Starling in Southern Africa using a species distribution modelling approach. Two different techniques were used which resulted in two potential distributions for the starling that were implemented in Objective II.

Objective II: I reconstructed the European Starling's range expansion using a dynamical modelling approach. Both potential distributions from Objective I as well as several other parameters were implemented in an individual based modelling framework designed for the European Starling in Southern Africa. Various combinations of these parameters lead to a total of 108 different models that were evaluated using two different independent data sets and applying numerous evaluation criteria. Three models that best fitted the evaluation criteria were chosen as representatives of the starling's range expansion process in Southern Africa.

Objective III: I forecasted future distributions of the European Starling in Southern Africa using the three best fitted models from Objective II. From these final distributions I found an optimal model that opened new doors for studying the starling's dispersal route and distributional dynamics in greater detail.

The hybrid model provided an opportunity to incorporate a variety of environmental features, behavioural dynamics, dispersal scenarios and geographic features that lead to a well-rounded study of the distribution dynamics of the European Starling in Southern Africa. Conducting parameter sensitivity analysis and finding an optimal model assisted the discovery of important environmental and behavioural features that drove the starling's range expansion and current distribution.

According to the optimal model, for the starlings to reach their observed distribution the flocks had 300 m elevation access between any two neighbouring grids (grid size of 10 km x 10 km), 200 km maximum dispersal distance per year and the cognitive ability to choose among five different locations the one with the highest habitat suitability. The model predicted that the starlings were able to manoeuvre around mountainous regions and avoided the semi-desert regions of the Karoo. The presence of the starling's cognitive ability is apparent in their dispersal decision when choosing among different locations. This supports the study by Tobler and Smith (2004) that certain starlings in the population act as floaters that will go explore first and gather information before settling in a new location.

According to the principle of parsimony, simpler models are often preferred above more complex models (Jongejans *et al*., 2008). In my model, I observed that when I added an additional distance factor in the dispersal function, thereby increasing the complexity of the starlings' cognitive ability, model performance decreased. Adding this additional level of complexity to the model therefore did not further contribute to our understanding of the starling's dynamics.

The complexity and biological realism of dynamical spatially explicit individual based models allow for intensive studies of the spatial dynamics of species within a landscape as compared to analytical models (Jongejans *et al*., 2008). Analytical models possess a greater ability of exploring model dynamics and parameters analytically, but lack the ability to incorporate as much biological and geographical realism as IBMs. IBMs are therefore still recommended whenever detailed information and sufficient data are available regarding the study species and the mechanisms of the system. In this study I presented a method of parameter sensitivity analysis for exploring the model dynamics incorporated in an IBM. Such methods could add to the current knowledge used for analysing IBMs more efficiently.

The European Starling, being such a versatile species, is highly complex to study. Nonetheless I was able to develop a model that effectively predicted their range expansion and when predicting future expansions an important observation was made: the starlings have not yet reached a saturation phase and are continuing to expand their range. This could pose potential problems for Southern Africa's future:

- some of our native bird species could be outcompeted by this resourceful bird and become endangered as was the case with the Australian parrots (Section 2.2.2),
- we could experience great economic loss, similar to what happened in North America (Section 2.2.2) if the starling's flock sizes becomes large enough to cause damage to agricultural lands, and/or
- we could experience health risks as in North America (Section 2.2.2) with the increase of flock sizes in developed areas.

A hybrid model, such as the one developed in this study, could be effectively implemented for studying newly introduced species and predicting their future dispersal routes. After such studies management plans could be structured accordingly for controlling the species' spread. Unfortunately, the European Starling has already progressed so far in their invasion process and are expanding rapidly, making them extremely difficult to manage. This study contributed to a greater awareness of the risks behind introducing species to new environments and if not controlled in their initial phases, they could disperse through an entire region, making any further spread uncontainable.

There lies great potential within dynamic spatial models. They have a lot to contribute to our knowledge of species, their distributions, their distribution dynamics and how to conserve and manage them in a changing environment.

5.2 Future Recommendations

The potential distribution map incorporated in the dynamic IBM played a crucial role in defining how the species will disperse. Chapter 3 could therefore be expanded to ensure that the best habitat suitability map is obtained for further studies. A detailed comparative study among various SDM methods, incorporating additional predictor variables, studying different background samples and sample sizes are necessary.

A number of simulations are required for each model in order to obtain an average performance when using an IBM. Due to insufficient computer power and limited time it was not possible to conduct multiple simulations for this study. A parameter sensitivity analysis could also be performed on the demographic parameters that were obtained from Hui *et al.* (2012), which were implemented in the IBM as constants. This could be important for validating the available knowledge of the starling's life-history parameters in Southern Africa.

I incorporated the most important biotic interaction for the starlings: humans. It may however be important to account for other biotic interactions as well. Further model development could incorporate species' genetics and spatial sorting mechanisms. Moving towards a mechanistic modelling approach might be necessary as a detailed study of the species' physiology will assist in determining the true limiting factors to their distribution (Carey, 1996).

A final cautionary note to take into consideration: the performance of spatially explicit models such as IBMs is influenced by spatial scale. The ecological processes incorporated into an IBM therefore require greater knowledge of the appropriate scales at which they should be studied.

BIBLIOGRAPHY

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223–1232.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33, 1677–1688.

Araújo, M.B., Pearson, R.G., Thuiller, W, Erhard, M., 2005. Validation of species – climate impact models under climate change. Global Change Biology 11, 1504–1513.

Araújo, M.B., Thuiller, W, Yoccoz, N.G., 2009. Reopening the climate envelope reveals macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences of the United States of America 106, E45–E46.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157, 101–118.

Austin, M.P., 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200, 1–19.

Bahn, V., McGill, B.J., 2013. Testing the predictive performance of distribution models. Oikos 122, 321–331.

Beale, C.M., Lennon, J.J., Gimona, A., 2008. Opening the climate envelope reveals no macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences of the United States of America 105, 14908–14912.

Berthouly-Salazar, C., Hui, C., Blackburn, T.M., Gaboriaud, C., Van Rensburg, B.J., Van Vuuren, B.J., Le Roux, J.J., 2013. Long-distance dispersal maximizes evolutionary potential during rapid geographic range expansion. Molecular ecology, in Press.

Bohonak, A.J., Van der Linde, K., 2004. RMA: Software for Reduced Major Axis Regression.

Brotons, L., Thuiller, W., Araújo, M.B., Hirzel, A.H., 2004. Presence-Absence versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability. Ecography 27, 437–448.

Cabe, P.R., 1999. Dispersal and Population Structure in the European Starling. The Condor 101, 451–454.

Carey, P.D., 1996. DISPERSE: A cellular automaton for predicting the distribution of species in a changed climate. Global Ecology and Biogeography 5, 217–226.

Caswell, H., Lensink, R., Neubert, M.G., Lensink, R.O.B., 2003. Demography and Dispersal: Life Table Response Experiments for Invasion Speed. Ecology 84, 1968–1978.

Chase, J.M., Leibold, M.A., 2003. Ecological Niches: Linking Classical and Contemporary Approaches. The University of Chicago Press, Chicago.

Chow, J., 2000. Sturnus Vulgaris. Animal Diversity Web. URL http://animaldiversity.ummz.umich.edu/site/accounts/information/Sturnus_vulgaris.html (accessed 6.10.12).

CIESIN, 2005. Wildlife Conservation Society (WCS), and Center for International Earth Science Information Network (CIESIN)/Columbia University. Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (Geographic). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). URL http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic (accessed 8.10.12).

Craig, A.J.F.K., 1997. European Starling. In: Harrison, J.A., Allan, D.G., Underhill, L.G., Herremans, M., Tree, A.J., Parker, V., Brown, C.J. (Eds.), The Atlas of Southern African Birds. BirdLife South Africa, Johannesburg, pp. 452–453.

Craney, T.A., Surles, J.G., 2002. Model-Dependent Variance Inflation Factor Cutoff Values. Quality Engineering 14, 391–403.

DeAngelis, D.L., Mooij, W.M., 2005. Individual-Based Modeling of Ecological and Evolutionary Processes. Annual Review of Ecology, Evolution, and Systematics 36, 147–168.

Elith, J., Ferrier, S., Guisan, A., Graham, C.H., Anderson, R.P., Dudı, M., Hijmans, R J, Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-pereira, R., Schapire, R.E., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Graham, C.H., 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1, 330–342.

Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40, 677–697.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. The Journal of Animal Ecology 77, 802–13.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43–57.

ESRI, 2011. ArcGIS Desktop: Release 10.

Fischl, J., Caccamise, D.F., 1985. Influence of Habitat and Season on Foraging Flock Composition in the European Starling (Sturnus Vulgaris). International Association for Ecology 67, 532–539.

Fischl, J., Caccamise, D.F., 1987. Relationships of Diet and Roosting Behavior in the European Starling. American Midland Naturalist 117, 395–404.

Franklin, J., 2010. Moving beyond static species distribution models in support of conservation biogeography. Diversity and Distributions 16, 321–330.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I., Thuiller, Wilfried, 2010. Predicting potential distributions of invasive species: where to go from here? Diversity and Distributions 16, 331–342.

Ghianni, T., 2013. Bird invasion brings real-life horror to Kentucky city. Reuters.

GLC2000, 2003. The Global Land Cover Map for the Year 2000. European Commision Joint Research Centre. URL http://www-gem.jrc.it/glc2000 (accessed 8.10.12).

Grimm, V., 1999. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? Ecological Modelling 115, 129–148.

Guisan, A., Thuiller, W, 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993–1009.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Harrison, J.A., Allan, D.G., Underhill, L.G., Herremans, M., Tree, A.J., Parker, V., Brown, C.J. (Eds.), 1997. The Atlas of Southern African Birds. Vol. 1: Non-passerines. BirdLife South Africa, Johannesburg.

Harrison, J.A., Underhill, L.G., Barnard, P., 2008. The seminal legacy of the Southern African Bird Atlas Project. South African Journal of Science 104, 82–84.

Hawthorne, B.L., 2012. Geospatial Modelling Environment. Spatial Ecology LLC. URL www.spatialecology.com (accessed 8.12.13).

Higgins, S.I., Richardson, D.M., Cowling, R.M., 1996a. Modeling invasive plant spread: the role of plant-environment interactions and model structure. Ecology 77, 2043–2054.

Higgins, S.I., Richardson, D.M., Cowling, R.M., Ecology, S., Oct, N., 1996b. Modeling Invasive Plant Spread : The Role of Plant-Environment Interactions and Model Structure. Ecology 77, 2043–2054.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978.

Hugueny, B., 1989. West African rivers as biogeographic islands: species richness of fish communities. Oecologia 79, 236–243.

Hui, C., Krug, R.M., Richardson, D.M., 2011. Modelling Spread in Invasion Ecology: A Synthesis. In: Richardson, D.M. (Ed.), Fifty Years of Invasion Ecology: The Legacy of Charles Elton. Blackwell Publishing Ltd, pp. 329–344.

Hui, C., Roura-pascual, N., Brotons, Lluís, Robinson, R.A., Evans, K.L., 2012. Flexible dispersal strategies in native and non-native ranges: environmental quality and the "good – stay, bad – disperse" rule. Ecography 35, 1024–1032.

Hui, C., Terblanche, J.S., Chown, S.L., McGeoch, M. a., 2010. Parameter landscapes unveil the bias in allometric prediction. Methods in Ecology and Evolution 1, 69–74.

Huntley, B., Collingham, Y.C., Willis, S.G., Green, R.E., 2008. Potential impacts of climatic change on European breeding birds. PloS one 3, e1439.

Hutchinson, G.E., 1957. Concluding Remarks. In: Cold Spring Harbor Symposia on Quantitative Biology. Connecticut, pp. 415–427.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the globe Version 4. CGIAR-CSI SRTM 90m Database. URL http://srtm.csi.cgiar.org (accessed 8.22.12).

Jongejans, E., Skarpaas, O., Shea, K., 2008. Dispersal, demography and spatial population models for conservation and control management. Perspectives in Plant Ecology, Evolution and Systematics 9, 153–170.

Judson, O.P., 1994. The rise of the individual-based model in ecology. TREE 9, 9–14.

Kearney, M., Porter, W., 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology letters 12, 334–350.

Kessel, B., 1953. Distribution and Migration of the European Starling in North America. The Condor 55, 49–67.

Kessel, B., 1957. A Study of the Breeding Biology of the European Starling (Sturnus vulgaris L.) in North America. American Midland Naturalist 58, 257–331.

Kirkpatrick, W., Woolnough, A., 2007. Common Starling. Government of Western Australia. Department of Agriculture and Food. Note: 253, Forrestfield.

Link, R., 2004. Living with Wildlife: Starlings. Washington Department of Fish and Wildlife. URL http://wdfw.wa.gov/living/starlings.html (accessed 7.16.13).

Linz, G.M., Homan, J.H., Gaukler, S.M., Penry, L.B., Bleier, W.J., 2007. European Starlings: A review of an invasive species with far-reaching impacts. In: Managing Vertebrate Invasive Species. pp. 378–386.

Liu, C., White, M., Newell, G., 2009. Measuring the accuracy of species distribution models: a review. In: 18th World IMACS / MODSIM Congress. pp. 4241–4247.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17, 145–151.

Long, J.L., Tingay, S., 1981. Introduced birds of the world: The worldwide history, distribution, and influence of birds introduced to new environments. Reed.

Lowe, S., Browne, M., Boudjelas, S., De Poorter, M., 2000. 100 of the World's Worst Invasive Alien Species A selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG).

Lynch, J.A., Messmer, T.A., 2010. European Starlings, Wildlife Damage Management Series. Utah State University Extension Service and College of Natural Resources.

Manel, S., Williams, C.H., Ormerod, S.J., 2001. Evaluating presence – absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.

Nakazato, T., Warren, D.L., Moyle, L.C., 2010. Ecological and geographic modes of species divergence in wild tomatoes. American Journal of Botany 97, 680–693.

Nehrbass, N., Winkler, E., 2007. Is the Giant Hogweed still a threat? An individual-based modelling approach for local invasion dynamics of Heracleum mantegazzianum. Ecological Modelling 201, 377–384.

O'brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity 41, 673–690.

Pearson, R.G., Dawson, T.P., Liu, C., 2004. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography 27, 285–298.

Phillips, B.S., 2008. A Brief Tutorial on Maxent.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259.

Phillips, S.J., Dudık, M., 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Pimentel, D., Lach, L., Zuniga, R., Morrison, D., 2000. Environmental and Economic Costs of Nonindigenous Species in the United States. BioScience 50, 53–65.

Pittman, S.J., Brown, K.A., 2011. Multi-scale approach for predicting fish species distributions across coral reef seascapes. PloS one 6, e20583.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (accessed 4.4.12).

Ridgeway, G., 2013. gbm: Generalized Boosted Regression Models. R package version 2.1. URL http://CRAN.R-project.org/package=gbm (accessed 4.4.12).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. URL http://www.biomedcentral.com/1471-2105/12/77/ (accessed 4.4.12).

Schuster, J., 2010. European Starlings Population Control Methods. Wild Wing Company. URL http://www.wildwingco.com/starlings.html (accessed 9.5.12).

Shigesada, N., Kawasaki, K., Takeda, Y., 1995. Modeling Stratified Diffusion in Biological Invasions. The American naturalist 146, 229–251.

Sinclair, I., Davidson, I., 2006. Southern African Birds: A Photographic Guide, Second. ed. Struik Publishers.

Sol, D., Timmermans, S., Lefebvre, L., 2002. Behavioural flexibility and invasion success in birds. Animal Behaviour 63, 495–502.

Stine, R.A., 1995. Graphical Interpretation of Variance Inflation Factors. The American Statistician 49, 53–56.

Thuiller, W, Brotons, L, Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172.

Tobler, M., Smith, H.G., 2004. Specific floater home ranges and prospective behaviour in the European starling , Sturnus vulgaris. Naturwissenschaften 91, 85–89.

Travis, J.M.J., Münkemüller, T., Burton, O.J., Best, A., Dytham, C., Johst, K., 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. Molecular Biology and Evolution 24, 2334–43.

Van den Bosch, F., Hengeveld, R., Metz, J.A.J., 1992. Analysing the velocity of animal range expansion. Journal of Biogeography 19, 135–150.

Vanderwal, J., Shoo, L.P., Graham, C., Williams, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220, 589–594.

Walsh, C., Mac Nally, R., 2013. hier.part: Hierarchical Partitioning. R package version 1.0-4. URL http://CRAN.R-project.org/package=hier.part(accessed 4.4.12).

Waterman, M., Fuller, C.O., Murray, M.D., 2008. Studies of roosting common starlings Sturnus vulgaris in South Australia. Corella 32, 25–29.

APPENDIX A: Variable Elimination Procedure for SDM2

Number of Variables		29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7
AUC (MaxEnt)		0.878	0.872	0.876	0.871	0.875	0.867	0.868	0.870	0.867	0.870	0.872	0.872	0.868	0.872	0.872	0.882	0.856	0.873	0.901	0.910	0.919	0.916	0.896
BIO1 Annual Mean Temperature	VIF	367.3	366.8	274	28.85	26.8	25.72	25.19	21.53	2.005	1.97	1.312	1.246	1.243	1.242	1.242								
	Percentage Contribution (MaxEnt)	1.3	1.3	1.2	1.5	1.4	1.7	1.6	1.6	1.7	1.4	1.6	1.7	1.6	1.5	1.6	1.5	3.6	3.6	3.8	8	30.4	35.5	36.6
	Relative Influence (BRT)	1.413	1.481	1.866	2.01	2.014	1.951	2.232	2.262	2.662	2.62	2.909	3.123	3.203	3.05	3.123	3.028	2.972	3.233	3.235	3.599	3.718	4.172	5.642
	Percentage Contribution (Hier part)																		4.779	4.283	0.765	0.575	0.528	0.542
BIO2 Mean Diurnal Range (Mean of monthly (max temp - min temp))	VIF	389.7	375.7	366	354.7																			
	Percentage Contribution (MaxEnt)	0.7	0.7	1.4	0.9																			
	Relative Influence (BRT)	0.745	0.971	0.997	1.062																			
	Percentage Contribution (Hier part)																							
BIO3 Isothermality (BIO2/BIO7) (* 100)	VIF	22.08	21.71	21.54	21.46	7.722	7.68	7.655	1.569	1.513	1.51	1.24												
	Percentage Contribution (MaxEnt)	0.9	0.9	0.8	0.9	0.7	0.8	0.4	1	0.9	1.3	0.9												
	Relative Influence (BRT)	0.281	0.271	0.399	0.284	0.337	0.324	0.357	0.526	0.542	0.37	0.485												
	Percentage Contribution (Hier part)																							

BIO4 Temperature Seasonality (standard deviation *100)	VIF	835.6	835	228.4	151.7	117.5	117.5	116.8														
	Percentage Contribution (MaxEnt)	1	1.4	1.2	1.1	1.2	1	1.1														
	Relative Influence (BRT)	0.916	1.394	1.467	1.443	1.693	1.538	1.782														
	Percentage Contribution (Hier part)																					
BIO5 Max Temperature of Warmest Month	VIF	2720	342	337.3	287.1	54.59	54.26	54.2	13.42	3.558	3.55											
	Percentage Contribution (MaxEnt)	1.1	1.1	1.3	1.8	1.5	1.4	1.2	1.2	1.2	1.2											
	Relative Influence (BRT)	0.837	0.866	0.96	1.054	1.061	1.105	1.251	1.309	1.603	1.47											
	Percentage Contribution (Hier part)																					
BIO6 Min Temperature of Coldest Month	VIF	4599	362.4	340.6	340.6	53.81	50.56	47.28	21.45													
	Percentage Contribution (MaxEnt)	0.5	0.5	0.5	0.4	0.6	0.5	0.6	0.4													
	Relative Influence (BRT)	1.607	1.848	2.037	1.953	2.077	2.167	2.064	2.258													
	Percentage Contribution (Hier part)																					
BIO7 Temperature Annual Range (BIO5-BIO6)	VIF	8782																				
	Percentage Contribution (MaxEnt)	0.8																				
	Relative Influence (BRT)	0.625																				
	Percentage Contribution (Hier part)																					

BIO10 Mean Temperature of Warmest Quarter	VIF	649.3	646.7	407.5																			
	Percentage Contribution (MaxEnt)	1.1	0.7	1.2																			
	Relative Influence (BRT)	0.53	0.66	0.635																			
	Percentage Contribution (Hier part)																						
BIO11 Mean Temperature of Coldest Quarter	VIF	1311	1311																				
	Percentage Contribution (MaxEnt)	0.4	0.4																				
	Relative Influence (BRT)	1.042	1.325																				
	Percentage Contribution (Hier part)																						
BIO12 Annual Precipitation	VIF	192.3	192.1	184	177.1	167.5	158.3																
	Percentage Contribution (MaxEnt)	0.6	0.7	0.6	0.7	0.6	0.8																
	Relative Influence (BRT)	0.936	1.054	1.147	1.07	1.198	1.455																
	Percentage Contribution (Hier part)																						
BIO15 Precipitation Seasonality (Coefficient of Variation)	VIF	13.17	13.16	13.14	12.25	12.13	10.5	10.26	9.141	8.779	1.57	1.255	1.212										
	Percentage Contribution (MaxEnt)	0.5	0.5	0.5	0.7	0.5	0.7	0.4	0.6	0.6	0.9	1	0.9										
	Relative Influence (BRT)	1.108	1.199	1.227	1.24	1.199	1.258	1.492	1.477	1.579	1.65	1.907	2.135										
	(Hier part)																						

BIO16 Precipitation of Wettest Quarter	VIF	207	206.4	206.2	199.3	194.8																
	Percentage Contribution (MaxEnt)	1.4	1.3	1.5	1.5	1.6																
	Relative Influence (BRT)	1.628	2.015	2.053	2.129	2.227																
	Percentage Contribution (Hier part)																					
BIO17 Precipitation of Driest Quarter	VIF	29.6	29.52	29.5	29.34	27.64	13.25	12.27	11.85	11.32												
	Percentage Contribution (MaxEnt)	0.9	0.9	0.9	0.9	0.9	1	0.8	1	1												
	Relative Influence (BRT)	1.068	1.008	1.055	1.126	1.267	1.16	1.417	1.231	1.351												
	Percentage Contribution (Hier part)																					
BIO18 Precipitation of Warmest Quarter	VIF	397	395.2	385.4	361.9	323.9	181.7	9.693	9.489	6.559	2.97	1.756	1.707	1.664	1.662	1.658						
	Percentage Contribution (MaxEnt)	4.9	4.3	4.7	4.9	4.3	6.1	6.2	5.5	5.9	6.2	6.4	4.8	5.1	4.3	5.2	6.2	6	6.3	6.7	5.3	2.2	1.9	1.7
	Relative Influence (BRT)	4.96	4.661	4.517	4.831	4.697	5.557	6.093	6.422	6.456	6.61	6.848	6.957	7.326	7.336	7.366	7.503	7.379	7.727	7.814	7.908	8.049	9.176	13.3
	Percentage Contribution (Hier part)																		11.01	14.09	14.55	13.02	13.82	12.92
BIO19 Precipitation of Coldest Quarter	VIF	189.7	188.7	182.2	168.8	151.5	66.22	6.485	6.284	6.03	2.36	1.717	1.717	1.701	1.697	1.696								
	Percentage Contribution (MaxEnt)	2.4	2.7	2.5	2.3	3	2.5	2.8	2.6	2.6	3.2	3.3	3.1	3.2	4.5	4.9	7	5.6	4.2	9.2	9.7	3.9	2.3	2.4
	Relative Influence (BRT)	25.21	24.54	24.32	24.41	24.47	25.17	24.65	25.01	25.53	26.8	26.33	25.96	26.38	26.98	26.96	27.56	27.76	27.69	28.97	29.56	30.97	31.74	37.07
	Percentage Contribution (Hier part)																		22.4	27.04	32.47	30.92	31.07	43.04

Human Footprint	VIF	1.415	1.414	1.406	1.405	1.402	1.397	1.36	1.352	1.313	1.31	1.307	1.299	1.271	1.271	1.27							
	Percentage Contribution (MaxEnt)	1.3	1.2	1.2	1.4	1.2	1.2	1.2	1.2	1.4	1.2	1.4	1.2	1.2	1.1	1.3	5.4	7.3	7.7	4.4	5.9	44.9	53.6	55.7
	Relative Influence (BRT)	4.022	4.044	4.172	4.123	4.209	4.258	4.198	4.251	4.395	4.46	4.435	4.669	4.836	4.744	4.711	4.681	4.704	4.791	4.668	5.239	5.435	5.692	8.514
	Percentage Contribution (Hier part)																		6.76	6.604	4.041	4.667	4.844	7.333
Closed evergreen lowland forest	VIF	1.288	1.288	1.288	1.253	1.253	1.239	1.239	1.238	1.204	1.2	1.203	1.196	1.166	1.14	1.139								
	Percentage Contribution (MaxEnt)	4.5	4.4	4.3	4.3	4.6	4.6	4.6	4.8	4.8	4.7	5	5.1	5	5.2	5	4.9							
	Relative Influence (BRT)	1.025	1.087	1.076	1.166	1.091	1.072	1.087	1.116	1.262	1.06	1.187	1.249	1.374	1.325	1.381	1.269							
	Percentage Contribution (Hier part)																							
Mosaic Forest / Savanna	VIF	1.234	1.233	1.233	1.225	1.222	1.217	1.21	1.204	1.203	1.2	1.199	1.197	1.197										
	Percentage Contribution (MaxEnt)	0.5	0.4	0.5	0.5	0.4	0.5	0.4	0.4	0.4	0.3	0.3	0.4	0.4										
	Relative Influence (BRT)	0.479	0.62	0.611	0.584	0.639	0.537	0.619	0.587	0.695	0.62	0.624	0.74	0.738										
	Percentage Contribution (Hier part)																							
Closed deciduous forest	VIF	1.444	1.443	1.441	1.438	1.438	1.437	1.435	1.434	1.434	1.43	1.414	1.406	1.372	1.292	1.28								
	Percentage Contribution (MaxEnt)	6.4	6.6	6.6	6.7	7.1	6.3	6.8	7	6.7	6.5	6.4	7	7.1	7.5	7.6	7	12.3	17.3	16.5	14.8	3.3		
	Relative Influence (BRT)	3.804	3.692	3.85	3.881	3.847	3.875	4.08	4.163	4.113	4.1	4.291	4.435	4.513	4.437	4.476	4.373	4.365	4.952	4.962	5.161	5.002		
	Percentage Contribution (Hier part)																		4.641	4.153	0.817	1.405		

Deciduous woodland	VIF	1.337	1.326	1.32	1.31	1.3	1.297	1.295	1.287	1.275	1.26	1.251	1.242	1.223	1.221	1.22							
	Percentage Contribution (MaxEnt)	10.5	10.8	10.4	10.7	10.8	10.7	10.7	11.9	10.8	10.6	10.3	11.3	11.1	11.9	11.8	12.1	14.7	16.1	16.8	13.2	4.8	1.9	1.1
	Relative Influence (BRT)	9.157	9.35	9.235	9.099	9.089	9.593	9.493	9.672	9.727	9.71	10.13	9.579	9.816	9.891	9.875	10.21	10.5	10.17	10.33	10.63	11.28	12.9	11.5
	Percentage Contribution (Hier part)																		5.114	4.757	2.07	3.772	4.282	4.429
Deciduous shrubland with sparse trees	VIF	1.212	1.212	1.211	1.205	1.205	1.205	1.179	1.176	1.174	1.17	1.169	1.16	1.159	1.157	1.157								
	Percentage Contribution (MaxEnt)	4.8	4.9	5	5	4.8	4.7	4.7	4.9	5.2	4.8	4.8	5	5	5	5.1	4.3	4.5						
	Relative Influence (BRT)	0.841	0.971	1.04	0.915	1.007	0.901	0.988	0.957	1.081	0.92	0.988	1.076	1.229	1.055	1.212	1.071							
	Percentage Contribution (Hier part)																							
Open deciduous shrubland	VIF	1.273	1.269	1.269	1.262	1.262	1.259	1.258	1.258	1.243	1.23	1.219	1.219	1.219	1.216	1.215								
	Percentage Contribution (MaxEnt)	8.1	8.6	8.5	8.3	9	8.9	9.1	9.3	9.3	9.3	9	9.8	9.9	9.6	9.5	8.7	8.9	5.7	5.7	6.2			
	Relative Influence (BRT)	3.416	3.556	3.637	3.528	3.739	3.51	3.934	3.938	4.281	3.87	4.027	4.371	4.547	4.184	4.42	4.039	4.13	4.425	4.038	4.07			
	Percentage Contribution (Hier part)																		4.574	4.093	0.812			
Closed grassland	VIF	1.153	1.152	1.151	1.151	1.151	1.138	1.124	1.122	1.122	1.12	1.114	1.085	1.084	1.084	1.084								
	Percentage Contribution (MaxEnt)	5.3	5	5.1	5.6	5.5	5.6	5.8	5.6	6.1	6.4	5.9	5.5	5.9	5.3	5.1	5.9	4.7	3.7					
	Relative Influence (BRT)	1.668	1.619	1.657	1.664	1.812	1.695	1.601	1.703	1.782	1.78	1.8	1.827	1.983	1.929	1.977	1.836	1.768	1.987					
	Percentage Contribution (Hier part)																		8.019					

Open grassland with sparse shrubs	VIF	1.31	1.307	1.305	1.295	1.295	1.292	1.273	1.27	1.266	1.26	1.255	1.254	1.236	1.236	1.235							
	Percentage Contribution (MaxEnt)	14.2	14.4	13.4	13	13.7	14.3	14.4	14.4	14.4	14.4	15.9	15.7	15.9	15.2	14.9	13.9	13.3	13.5	14.4	16.1	4.7	2.1	1.4
	Relative Influence (BRT)	10.05	9.895	10.33	10.25	10.1	10.43	10.33	10.52	10.49	10.7	10.9	10.89	10.68	11.34	10.87	11.09	11.01	11.12	11.32	11.68	12.48	12.72	12.53
	Percentage Contribution (Hier part)																		9.807	11.45	15.39	22.19	22.01	21.36
Open grassland	VIF	1.204	1.201	1.201	1.199	1.177	1.176	1.176	1.174	1.17	1.17	1.169	1.149	1.142	1.142	1.142								
	Percentage Contribution (MaxEnt)	5.6	5.6	6	5.9	5.8	6.1	5.7	5.7	5.9	6	6.3	5.9	5.8	6.5	6.4	6.7	4.8	3.5	2.6				
	Relative Influence (BRT)	2.366	2.543	2.565	2.533	2.66	2.469	2.55	2.566	2.677	2.6	2.737	2.736	2.765	2.705	2.729	2.784	2.816	2.878	2.569				
	Percentage Contribution (Hier part)																		6.226	6.652				
Sparse grassland	VIF	1.083	1.082	1.082	1.082	1.078	1.078	1.078	1.072	1.071	1.07	1.067	1.063	1.055	1.055	1.055								
	Percentage Contribution (MaxEnt)	5.3	5.5	5.5	5.3	5.7	5.2	5.3	5.5	5.6	5.8	5	5.5	5.4	6.1	6.2	6	4.9	3.8	3.8	4.7	1.3	0.9	
	Relative Influence (BRT)	10.51	10.01	10.03	10.16	10.14	10.32	10.1	10.17	10.1	10.8	10.44	10.44	10.45	10.78	10.67	10.72	10.96	10.86	11.48	11.41	12.11	12.13	
	Percentage Contribution (Hier part)																		9.43	9.97	10.54	13.7	13.75	
Croplands (>50%)	VIF	1.305	1.303	1.288	1.286	1.285	1.283	1.282	1.27	1.267	1.25	1.221	1.218	1.216	1.216	1.215								
	Percentage Contribution (MaxEnt)	10.3	10.7	10.4	10.5	10.6	10.5	11.1	10.7	10.8	10.7	11.3	12	12	12	11.4	10.2	9.5	14.6	16.3	16.1	4.4	1.9	1.2
	Relative Influence (BRT)	9.425	8.932	8.708	9.088	8.957	9.293	9.196	9.395	9.197	9.5	9.473	9.227	9.604	9.755	9.808	9.825	10.28	10.16	10.61	10.74	10.95	11.48	11.45
	Percentage Contribution (Hier part)																		7.238	6.903	18.54	9.758	9.69	10.37

Waterbodies	VIF	1.069	1.068	1.067	1.067	1.067	1.066	1.064	1.063	1.059	1.06	1.048	1.045	1.043	1.019							
	Percentage Contribution (MaxEnt)	0.8	0.8	1	1.2	0.9	0.9	0.9	0.9	0.9	1.1	0.8	0.8	0.8	0.6							
	Relative Influence (BRT)	0.032	0.057	0.032	0.052	0.06	0.056	0.049	0.075	0.056	0.06	0.051	0.081	0.072	0.057							
	Percentage Contribution (Hier part)																					
Cities	VIF	1.135	1.133	1.13	1.119	1.118	1.117	1.113	1.113	1.11	1.11	1.099	1.097	1.097	1.097	1.097						
	Percentage Contribution (MaxEnt)	3.8	3.5	3.8	3.6	3.7	4.2	4.3	3.9	3.9	3.8	4.4	4.4	4.6	3.6	4						
	Relative Influence (BRT)	0.304	0.332	0.376	0.352	0.406	0.304	0.433	0.381	0.419	0.31	0.432	0.506	0.471	0.442	0.423						
	Percentage Contribution (Hier part)																					

APPENDIX B: Reference Names for Model Simulations

SDM1 and SDM2 refer to the habitat suitability map obtained from the first and second species distribution modelling approach in Chapter 3. DEM100, DEM200 and DEM300 refer to elevation access of 100 m, 200 m or 300 m. CH2, CH5 and CH10 refer to the number of dispersal choices given to the starling amongst which the new locations were chosen. 5 to 100km, 5 to 200km and 5 to 300km refer to the dispersal distances applied in the dispersal kernel, varying with maximum distances of 100 km, 200 km and 300 km. HS and HS DIS refer to the decision function for choosing best new location; this function is based either on comparing their various habitat suitability values (only HS), or both habitat suitability and distance from current location (HS DIS).

Simulation Reference Name	Description
Sim1.1	SDM1, DEM100, CH2, 5 to 100km, HS
Sim1.2	SDM1, DEM100, CH2, 5 to 100km, HS DIS
Sim1.3	SDM1, DEM100, CH2, 5 to 200km, HS
Sim1.4	SDM1, DEM100, CH2, 5 to 200km, HS DIS
Sim1.5	SDM1, DEM100, CH2, 5 to 300km, HS
Sim1.6	SDM1, DEM100, CH2, 5 to 300km, HS DIS
Sim1.7	SDM1, DEM100, CH5, 5 to 100km, HS
Sim1.8	SDM1, DEM100, CH5, 5 to 100km, HS DIS
Sim1.9	SDM1, DEM100, CH5, 5 to 200km, HS
Sim1.10	SDM1, DEM100, CH5, 5 to 200km, HS DIS
Sim1.11	SDM1, DEM100, CH5, 5 to 300km, HS
Sim1.12	SDM1, DEM100, CH5, 5 to 300km, HS DIS
Sim1.13	SDM1, DEM100, CH10, 5 to 100km, HS
Sim1.14	SDM1, DEM100, CH10, 5 to 100km, HS DIS
Sim1.15	SDM1, DEM100, CH10, 5 to 200km, HS
Sim1.16	SDM1, DEM100, CH10, 5 to 200km, HS DIS
Sim1.17	SDM1, DEM100, CH10, 5 to 300km, HS
Sim1.18	SDM1, DEM100, CH10, 5 to 300km, HS DIS
Sim1.19	SDM1, DEM200, CH2, 5 to 100km, HS
Sim1.20	SDM1, DEM200, CH2, 5 to 100km, HS DIS
Sim1.21	SDM1, DEM200, CH2, 5 to 200km, HS
Sim1.22	SDM1, DEM200, CH2, 5 to 200km, HS DIS
Sim1.23	SDM1, DEM200, CH2, 5 to 300km, HS
Sim1.24	SDM1, DEM200, CH2, 5 to 300km, HS DIS
Sim1.25	SDM1, DEM200, CH5, 5 to 100km, HS
Sim1.26	SDM1, DEM200, CH5, 5 to 100km, HS DIS
Sim1.27	SDM1, DEM200, CH5, 5 to 200km, HS
Sim1.28	SDM1, DEM200, CH5, 5 to 200km, HS DIS
Sim1.29	SDM1, DEM200, CH5, 5 to 300km, HS
Sim1.30	SDM1, DEM200, CH5, 5 to 300km, HS DIS
Sim1.31	SDM1, DEM200, CH10, 5 to 100km, HS
Sim1.32	SDM1, DEM200, CH10, 5 to 100km, HS DIS
Sim1.33	SDM1, DEM200, CH10, 5 to 200km, HS
Sim1.34	SDM1, DEM200, CH10, 5 to 200km, HS DIS
Sim1.35	SDM1, DEM200, CH10, 5 to 300km, HS
Sim1.36	SDM1, DEM200, CH10, 5 to 300km, HS DIS
Sim1.37	SDM1, DEM300, CH2, 5 to 100km, HS
Sim1.38	SDM1, DEM300, CH2, 5 to 100km, HS DIS
Sim1.39	SDM1, DEM300, CH2, 5 to 200km, HS

Sim1.40	SDM1, DEM300, CH2, 5 to 200km, HS DIS
Sim1.41	SDM1, DEM300, CH2, 5 to 300km, HS
Sim1.42	SDM1, DEM300, CH2, 5 to 300km, HS DIS
Sim1.43	SDM1, DEM300, CH5, 5 to 100km, HS
Sim1.44	SDM1, DEM300, CH5, 5 to 100km, HS DIS
Sim1.45	SDM1, DEM300, CH5, 5 to 200km, HS
Sim1.46	SDM1, DEM300, CH5, 5 to 200km, HS DIS
Sim1.47	SDM1, DEM300, CH5, 5 to 300km, HS
Sim1.48	SDM1, DEM300, CH5, 5 to 300km, HS DIS
Sim1.49	SDM1, DEM300, CH10, 5 to 100km, HS
Sim1.50	SDM1, DEM300, CH10, 5 to 100km, HS DIS
Sim1.51	SDM1, DEM300, CH10, 5 to 200km, HS
Sim1.52	SDM1, DEM300, CH10, 5 to 200km, HS DIS
Sim1.53	SDM1, DEM300, CH10, 5 to 300km, HS
Sim1.54	SDM1, DEM300, CH10, 5 to 300km, HS DIS
Sim2.1	SDM2, DEM100, CH2, 5 to 100km, HS
Sim2.2	SDM2, DEM100, CH2, 5 to 100km, HS DIS
Sim2.3	SDM2, DEM100, CH2, 5 to 200km, HS
Sim2.4	SDM2, DEM100, CH2, 5 to 200km, HS DIS
Sim2.5	SDM2, DEM100, CH2, 5 to 300km, HS
Sim2.6	SDM2, DEM100, CH2, 5 to 300km, HS DIS
Sim2.7	SDM2, DEM100, CH5, 5 to 100km, HS
Sim2.8	SDM2, DEM100, CH5, 5 to 100km, HS DIS
Sim2.9	SDM2, DEM100, CH5, 5 to 200km, HS
Sim2.10	SDM2, DEM100, CH5, 5 to 200km, HS DIS
Sim2.11	SDM2, DEM100, CH5, 5 to 300km, HS
Sim2.12	SDM2, DEM100, CH5, 5 to 300km, HS DIS
Sim2.13	SDM2, DEM100, CH10, 5 to 100km, HS
Sim2.14	SDM2, DEM100, CH10, 5 to 100km, HS DIS
Sim2.15	SDM2, DEM100, CH10, 5 to 200km, HS
Sim2.16	SDM2, DEM100, CH10, 5 to 200km, HS DIS
Sim2.17	SDM2, DEM100, CH10, 5 to 300km, HS
Sim2.18	SDM2, DEM100, CH10, 5 to 300km, HS DIS
Sim2.19	SDM2, DEM200, CH2, 5 to 100km, HS
Sim2.20	SDM2, DEM200, CH2, 5 to 100km, HS DIS
Sim2.21	SDM2, DEM200, CH2, 5 to 200km, HS
Sim2.22	SDM2, DEM200, CH2, 5 to 200km, HS DIS
Sim2.23	SDM2, DEM200, CH2, 5 to 300km, HS
Sim2.24	SDM2, DEM200, CH2, 5 to 300km, HS DIS
Sim2.25	SDM2, DEM200, CH5, 5 to 100km, HS
Sim2.26	SDM2, DEM200, CH5, 5 to 100km, HS DIS
Sim2.27	SDM2, DEM200, CH5, 5 to 200km, HS
Sim2.28	SDM2, DEM200, CH5, 5 to 200km, HS DIS
Sim2.29	SDM2, DEM200, CH5, 5 to 300km, HS
Sim2.30	SDM2, DEM200, CH5, 5 to 300km, HS DIS
Sim2.31	SDM2, DEM200, CH10, 5 to 100km, HS
Sim2.32	SDM2, DEM200, CH10, 5 to 100km, HS DIS
Sim2.33	SDM2, DEM200, CH10, 5 to 200km, HS
Sim2.34	SDM2, DEM200, CH10, 5 to 200km, HS DIS
Sim2.35	SDM2, DEM200, CH10, 5 to 300km, HS
Sim2.36	SDM2, DEM200, CH10, 5 to 300km, HS DIS
Sim2.37	SDM2, DEM300, CH2, 5 to 100km, HS
Sim2.38	SDM2, DEM300, CH2, 5 to 100km, HS DIS
Sim2.39	SDM2, DEM300, CH2, 5 to 200km, HS
Sim2.40	SDM2, DEM300, CH2, 5 to 200km, HS DIS
Sim2.41	SDM2, DEM300, CH2, 5 to 300km, HS
Sim2.42	SDM2, DEM300, CH2, 5 to 300km, HS DIS
Sim2.43	SDM2, DEM300, CH5, 5 to 100km, HS
Sim2.44	SDM2, DEM300, CH5, 5 to 100km, HS DIS
Sim2.45	SDM2, DEM300, CH5, 5 to 200km, HS
Sim2.46	SDM2, DEM300, CH5, 5 to 200km, HS DIS

Sim2.47	SDM2, DEM300, CH5, 5 to 300km, HS
Sim2.48	SDM2, DEM300, CH5, 5 to 300km, HS DIS
Sim2.49	SDM2, DEM300, CH10, 5 to 100km, HS
Sim2.50	SDM2, DEM300, CH10, 5 to 100km, HS DIS
Sim2.51	SDM2, DEM300, CH10, 5 to 200km, HS
Sim2.52	SDM2, DEM300, CH10, 5 to 200km, HS DIS
Sim2.53	SDM2, DEM300, CH10, 5 to 300km, HS
Sim2.54	SDM2, DEM300, CH10, 5 to 300km, HS DIS
Null1.1	SDM1, DEM0, CH5, 5 to 200km, HS
Null1.2	SDM1, DEM200, CH1, 5 to 200km
Null1.3	SDM1, DEM0, CH1, 5 to 200km
Null2.1	SDM2, DEM0, CH5, 5 to 200km, HS
Null2.2	SDM2, DEM200, CH1, 5 to 200km
Null2.3	SDM2, DEM0, CH1, 5 to 200km

APPENDIX C: Historical Records' Evaluation Results

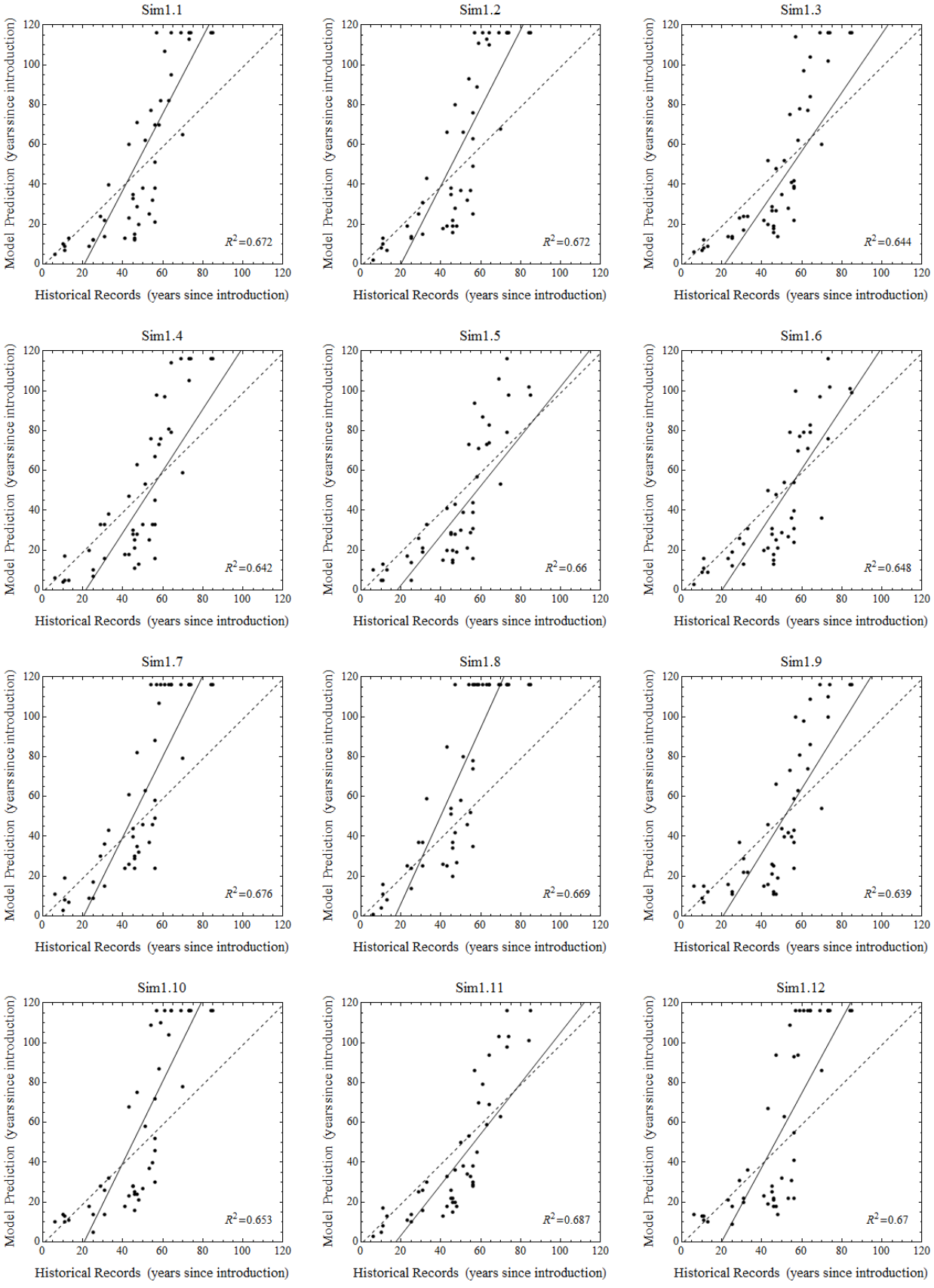
OLS – Ordinary least squares regression
RMA – Reduced major axis regression
SE – Standard error
R^2 – Coefficient of determination

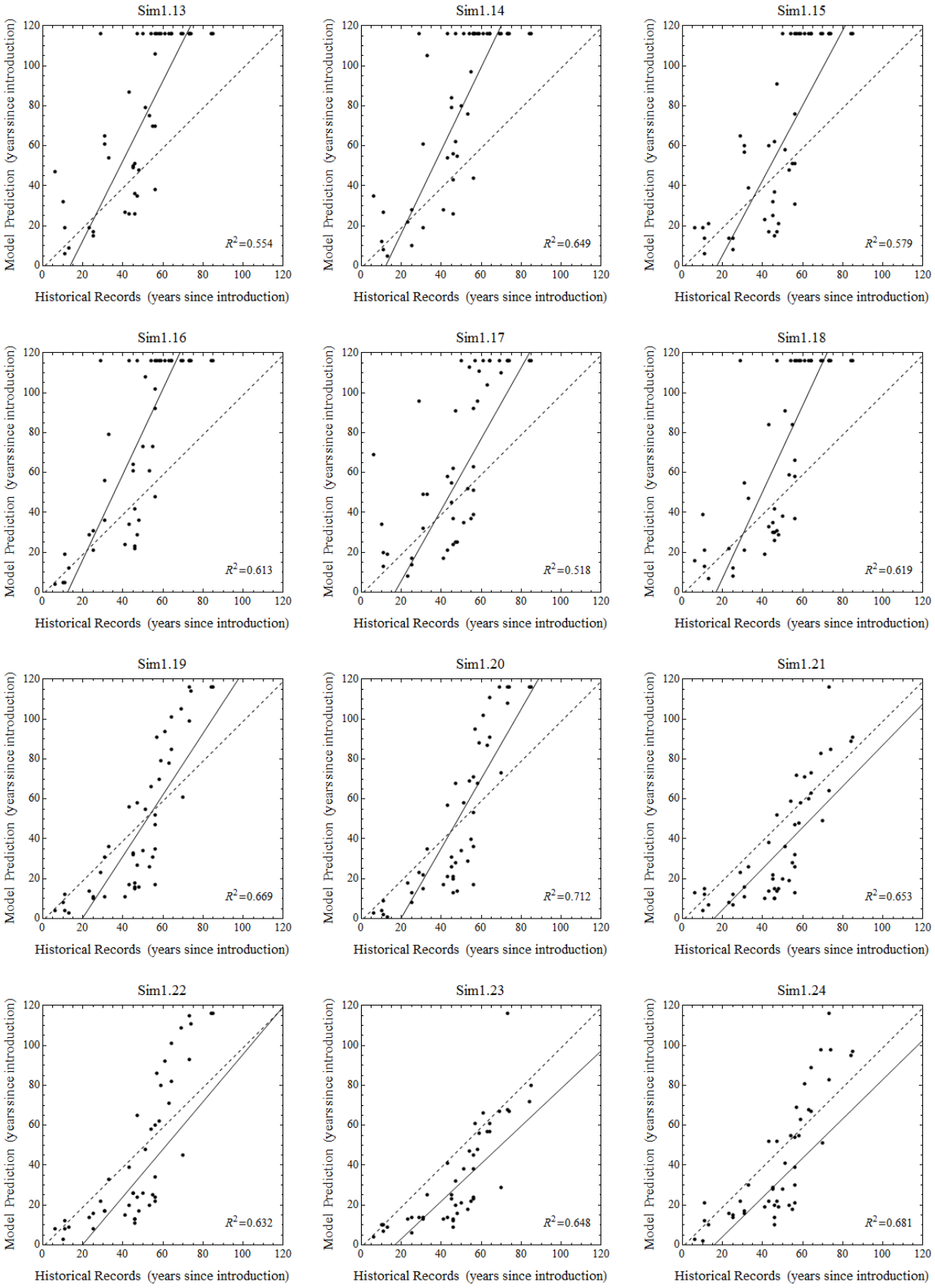
	Intercept			Slope			
Model	OLS	RMA	SE	OLS	RMA	SE	R^2
Sim1.1	-27.327	-45.020	9.030	1.646	2.018	0.176	0.665
Sim1.2	-26.476	-44.873	9.451	1.734	2.121	0.184	0.668
Sim1.3	-26.736	-42.684	8.345	1.566	1.901	0.163	0.678
Sim1.4	-25.764	-42.480	8.565	1.566	1.918	0.167	0.667
Sim1.5	-20.838	-35.722	7.481	1.336	1.650	0.146	0.656
Sim1.6	-18.140	-33.474	7.548	1.314	1.637	0.147	0.645
Sim1.7	-24.497	-41.733	9.200	1.768	2.131	0.179	0.689
Sim1.8	-19.657	-36.624	9.223	1.812	2.169	0.180	0.698
Sim1.9	-23.540	-40.729	8.553	1.509	1.871	0.167	0.651
Sim1.10	-26.493	-45.610	9.633	1.726	2.129	0.188	0.658
Sim1.11	-23.809	-37.296	7.227	1.396	1.680	0.141	0.691
Sim1.12	-25.208	-48.373	10.712	1.723	2.211	0.209	0.608
Sim1.13	0.521	-23.425	10.436	1.554	2.058	0.203	0.570
Sim1.14	2.721	-19.546	10.199	1.621	2.089	0.199	0.602
Sim1.15	-15.139	-38.333	10.661	1.701	2.190	0.208	0.604
Sim1.16	-6.249	-28.960	10.487	1.684	2.162	0.204	0.607
Sim1.17	-2.409	-29.642	10.914	1.450	2.023	0.213	0.514
Sim1.18	-9.538	-35.138	11.088	1.638	2.176	0.216	0.566
Sim1.19	-27.072	-41.678	7.928	1.555	1.862	0.155	0.697
Sim1.20	-30.252	-44.728	8.171	1.679	1.984	0.159	0.716
Sim1.21	-18.548	-33.191	6.954	1.157	1.465	0.136	0.624
Sim1.22	-26.358	-43.423	8.456	1.484	1.843	0.165	0.648
Sim1.23	-15.173	-27.641	6.020	1.023	1.285	0.117	0.633
Sim1.24	-18.860	-32.925	7.079	1.266	1.562	0.138	0.657
Sim1.25	-27.709	-42.090	7.916	1.579	1.881	0.154	0.704
Sim1.26	-23.284	-45.657	10.630	1.770	2.241	0.207	0.624
Sim1.27	-20.306	-34.897	7.174	1.247	1.554	0.140	0.644
Sim1.28	-30.054	-50.107	9.972	1.758	2.180	0.194	0.650
Sim1.29	-13.230	-26.435	6.472	1.121	1.398	0.126	0.642
Sim1.30	-27.618	-43.632	8.605	1.667	2.004	0.168	0.692
Sim1.31	-12.816	-31.344	9.436	1.713	2.103	0.184	0.664
Sim1.32	1.425	-19.162	9.785	1.630	2.063	0.191	0.624
Sim1.33	-9.728	-33.114	10.383	1.584	2.077	0.202	0.582
Sim1.34	-11.313	-30.458	9.584	1.704	2.107	0.187	0.654
Sim1.35	-14.866	-32.642	8.665	1.490	1.864	0.169	0.639
Sim1.36	-14.307	-38.403	10.793	1.666	2.173	0.210	0.588
Sim1.37	-25.427	-38.681	7.224	1.424	1.703	0.141	0.699

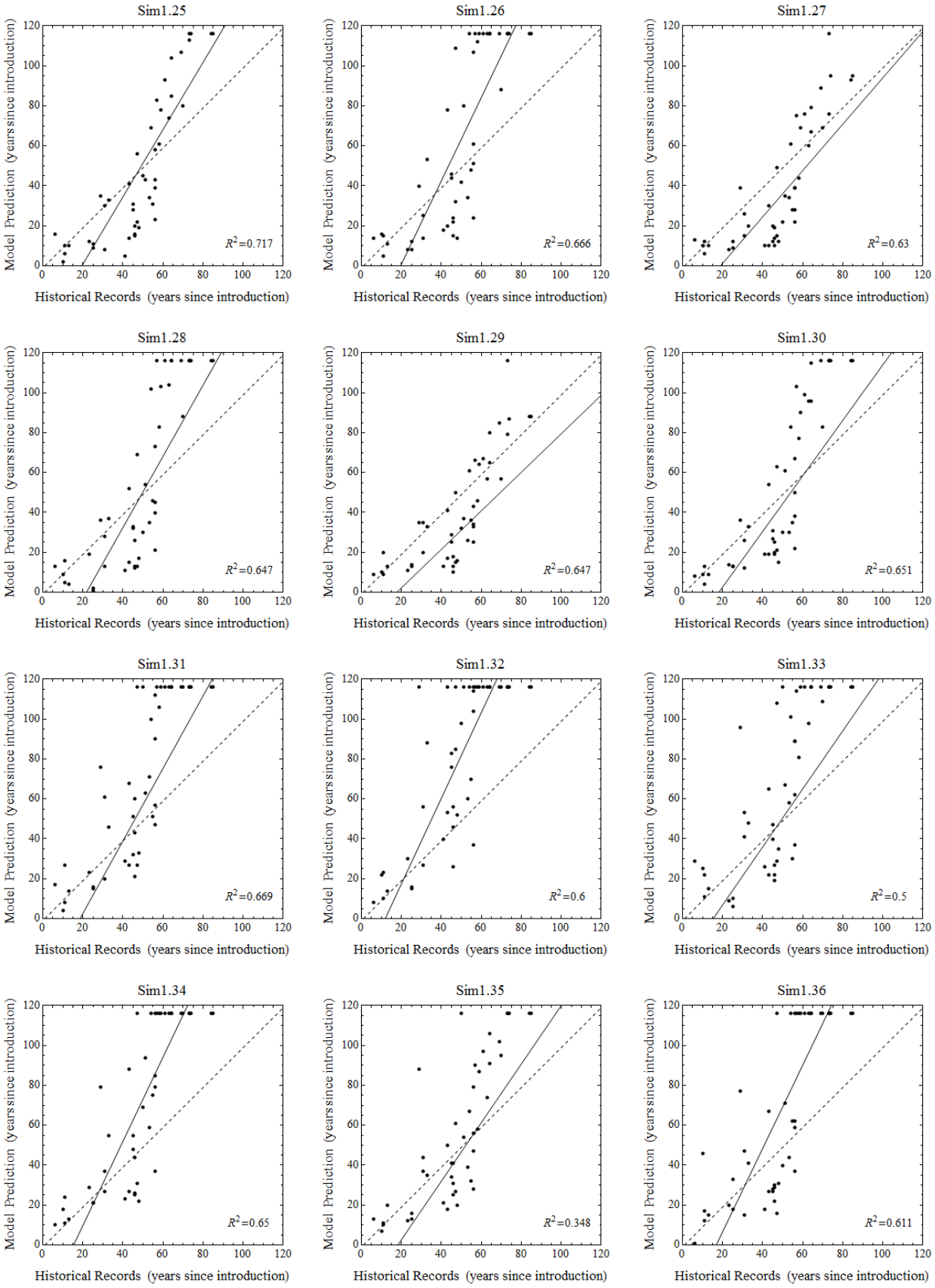
Sim1.38	-30.908	-48.202	8.865	1.623	1.986	0.173	0.667
Sim1.39	-11.642	-21.312	4.993	0.922	1.125	0.097	0.671
Sim1.40	-18.971	-31.846	6.413	1.133	1.404	0.125	0.651
Sim1.41	-10.514	-18.708	4.461	0.878	1.051	0.087	0.699
Sim1.42	-12.630	-26.135	6.119	0.959	1.243	0.119	0.595
Sim1.43	-26.183	-41.565	8.206	1.576	1.900	0.160	0.688
Sim1.44	-25.882	-46.253	10.279	1.845	2.274	0.200	0.658
Sim1.45	-17.653	-29.683	6.279	1.175	1.428	0.122	0.677
Sim1.46	-26.073	-41.694	8.186	1.539	1.868	0.160	0.679
Sim1.47	-16.542	-28.727	6.056	1.067	1.323	0.118	0.650
Sim1.48	-23.452	-39.173	8.215	1.539	1.870	0.160	0.677
Sim1.49	-13.302	-35.537	10.201	1.624	2.092	0.199	0.603
Sim1.50	-1.558	-23.848	10.353	1.675	2.144	0.202	0.610
Sim1.51	-9.986	-27.735	8.327	1.364	1.738	0.162	0.616
Sim1.52	-13.205	-31.577	9.290	1.672	2.058	0.181	0.660
Sim1.53	-10.800	-32.278	8.966	1.260	1.712	0.175	0.542
Sim1.54	-15.242	-36.940	10.439	1.766	2.222	0.203	0.631
Sim2.1	-23.464	-40.005	8.551	1.581	1.929	0.167	0.672
Sim2.2	-22.348	-39.138	8.689	1.609	1.962	0.169	0.672
Sim2.3	-17.601	-31.425	6.795	1.181	1.471	0.132	0.644
Sim2.4	-18.520	-33.194	7.186	1.243	1.552	0.140	0.642
Sim2.5	-11.371	-22.485	5.626	1.014	1.248	0.110	0.660
Sim2.6	-16.374	-30.506	6.996	1.227	1.524	0.136	0.648
Sim2.7	-24.126	-41.227	8.912	1.664	2.024	0.174	0.676
Sim2.8	-18.757	-37.837	9.820	1.806	2.207	0.191	0.669
Sim2.9	-17.889	-33.397	7.556	1.299	1.625	0.147	0.639
Sim2.10	-23.607	-42.362	9.365	1.659	2.054	0.182	0.653
Sim2.11	-11.706	-22.077	5.515	1.055	1.273	0.107	0.687
Sim2.12	-20.801	-36.861	8.276	1.524	1.862	0.161	0.670
Sim2.13	-2.960	-27.222	10.319	1.487	1.998	0.201	0.554
Sim2.14	-5.997	-25.208	9.529	1.675	2.079	0.186	0.649
Sim2.15	-10.779	-32.106	9.419	1.427	1.876	0.184	0.579
Sim2.16	-4.221	-26.277	10.295	1.676	2.140	0.201	0.613
Sim2.17	-6.027	-29.764	9.572	1.283	1.782	0.187	0.518
Sim2.18	-14.136	-36.104	10.344	1.703	2.165	0.202	0.619
Sim2.19	-16.972	-30.301	6.860	1.262	1.542	0.134	0.669
Sim2.20	-21.348	-34.412	7.248	1.473	1.745	0.141	0.712
Sim2.21	-6.713	-16.121	4.703	0.834	1.032	0.092	0.653
Sim2.22	-11.837	-23.457	5.596	0.948	1.192	0.109	0.632
Sim2.23	-6.933	-15.682	4.329	0.758	0.943	0.084	0.648
Sim2.24	-7.306	-15.502	4.308	0.813	0.985	0.084	0.681
Sim2.25	-20.524	-32.761	6.922	1.426	1.683	0.135	0.717
Sim2.26	-22.586	-40.786	9.316	1.702	2.085	0.182	0.666
Sim2.27	-10.517	-21.896	5.461	0.921	1.160	0.106	0.630
Sim2.28	-22.272	-38.882	8.211	1.437	1.786	0.160	0.647
Sim2.29	-8.105	-17.105	4.445	0.777	0.967	0.087	0.647
Sim2.30	-12.429	-25.216	6.369	1.125	1.394	0.124	0.651

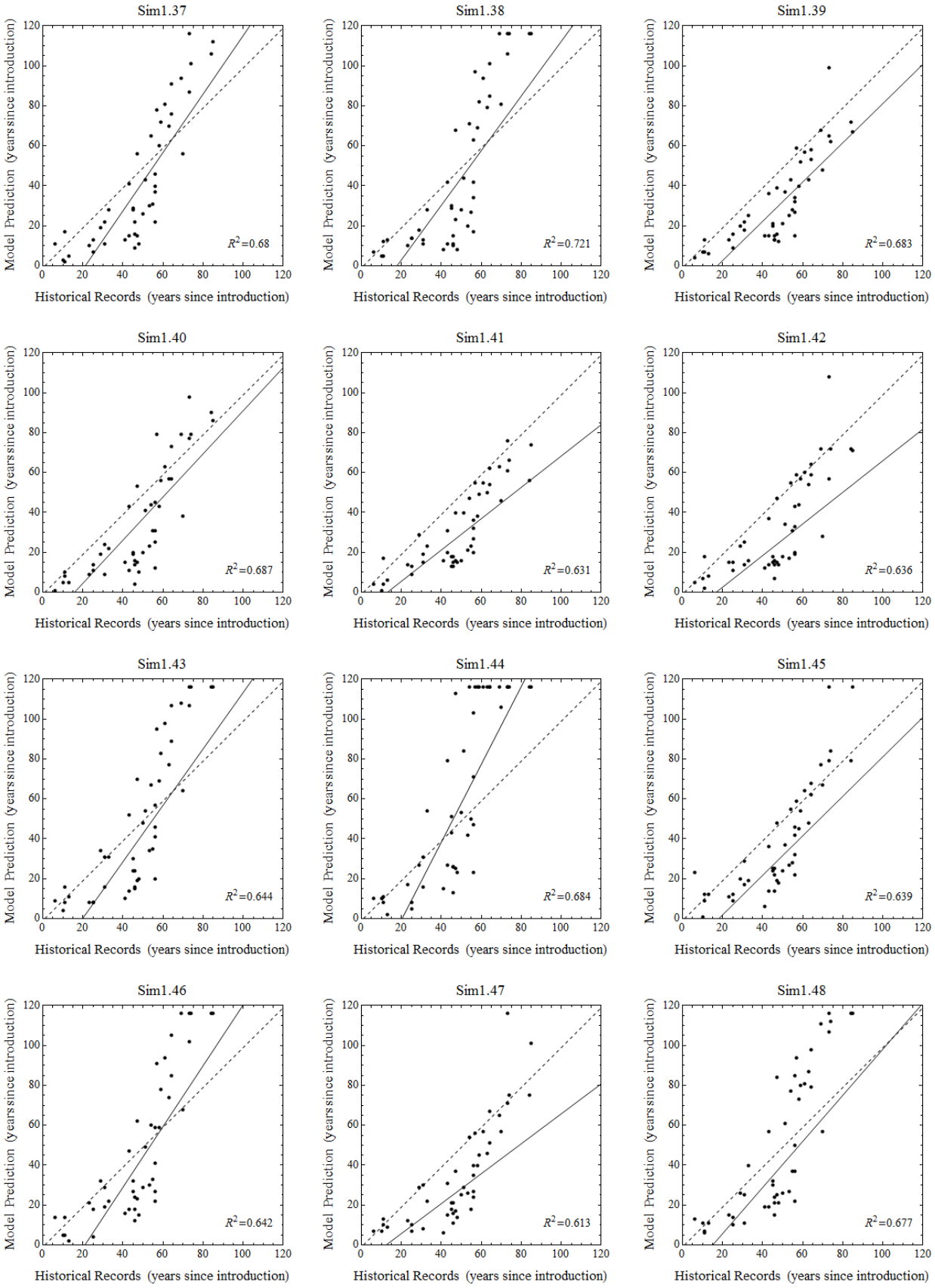
Sim2.31	-17.992	-33.732	8.098	1.488	1.820	0.158	0.669
Sim2.32	-2.498	-25.530	10.516	1.664	2.149	0.205	0.600
Sim2.33	-2.009	-22.272	7.965	1.030	1.456	0.155	0.500
Sim2.34	-13.015	-32.448	9.663	1.703	2.112	0.188	0.650
Sim2.35	1.917	-26.701	9.173	0.866	1.469	0.179	0.348
Sim2.36	-13.486	-35.201	10.093	1.634	2.091	0.197	0.611
Sim2.37	-18.728	-30.913	6.397	1.205	1.462	0.125	0.680
Sim2.38	-14.546	-24.330	5.581	1.161	1.367	0.109	0.721
Sim2.39	-8.884	-16.981	4.277	0.812	0.982	0.083	0.683
Sim2.40	-8.286	-17.096	4.686	0.897	1.082	0.091	0.687
Sim2.41	-2.500	-10.183	3.693	0.624	0.786	0.072	0.631
Sim2.42	-5.636	-13.280	3.704	0.632	0.793	0.072	0.636
Sim2.43	-14.738	-28.024	6.532	1.135	1.415	0.127	0.644
Sim2.44	-23.568	-39.633	8.493	1.614	1.952	0.166	0.684
Sim2.45	-8.455	-17.894	4.604	0.792	0.991	0.090	0.639
Sim2.46	-17.568	-31.984	7.060	1.221	1.524	0.138	0.642
Sim2.47	-1.511	-9.266	3.616	0.588	0.751	0.070	0.613
Sim2.48	-7.678	-17.407	5.082	0.952	1.156	0.099	0.677
Sim2.49	-5.770	-21.228	7.185	1.164	1.489	0.140	0.611
Sim2.50	-11.744	-30.358	9.542	1.746	2.138	0.186	0.667
Sim2.51	-1.083	-24.562	8.351	0.932	1.426	0.163	0.427
Sim2.52	-13.781	-41.524	11.692	1.665	2.248	0.228	0.548
Sim2.53	1.823	-21.294	7.434	0.706	1.192	0.145	0.351
Sim2.54	-16.179	-37.128	9.837	1.614	2.054	0.192	0.617

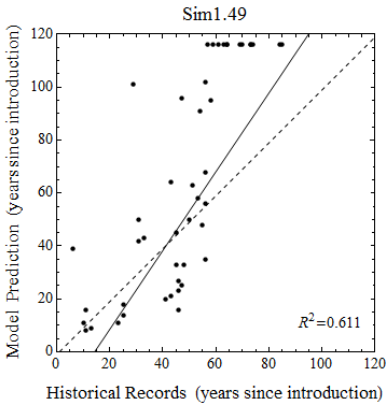
APPENDIX D: Reduced Major Axes Regression Model Fits

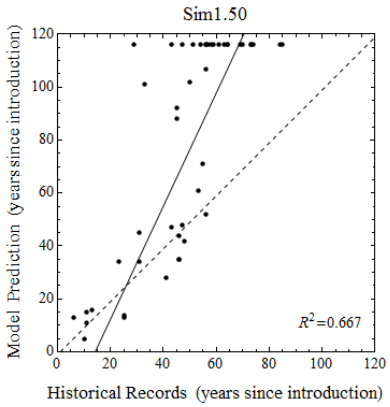


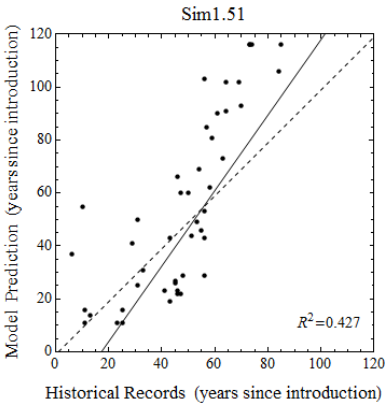


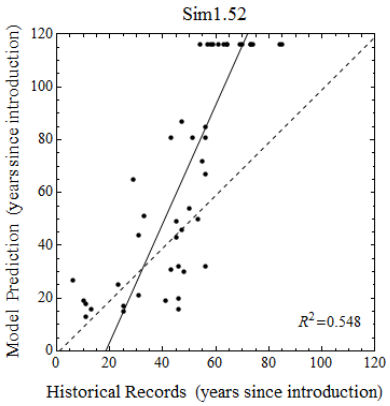


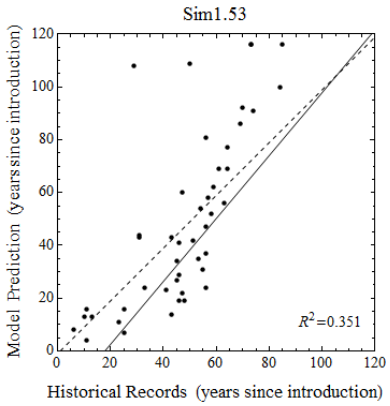


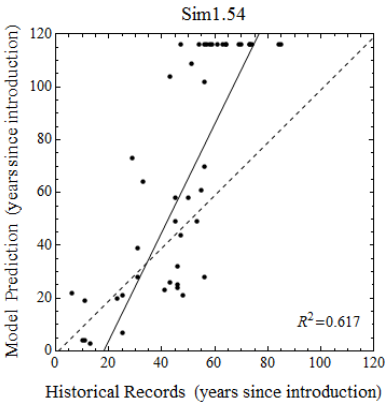


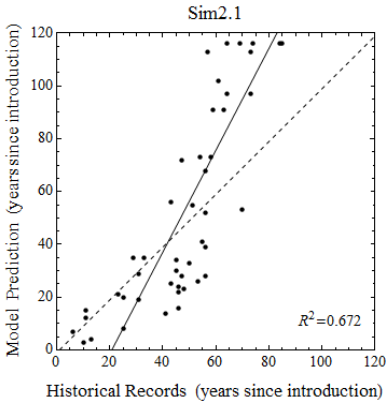


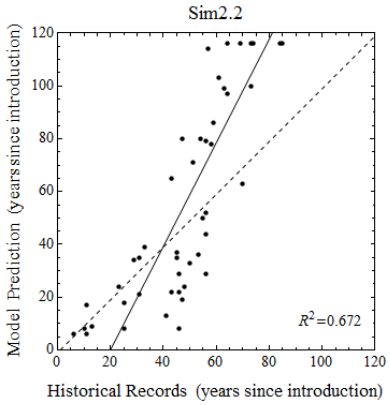


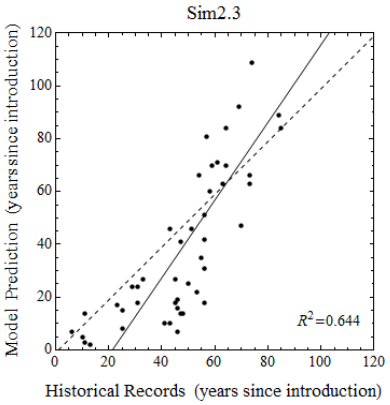


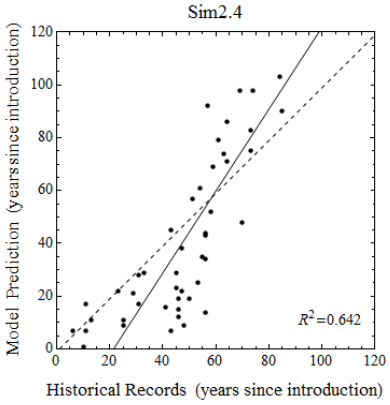


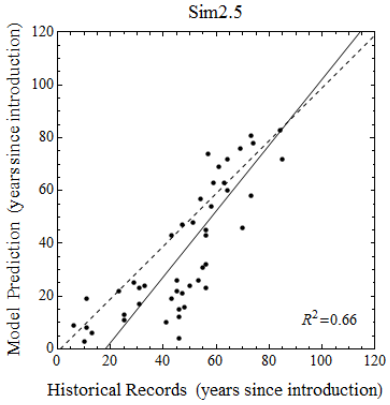


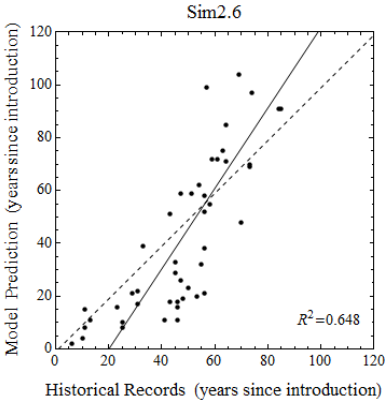


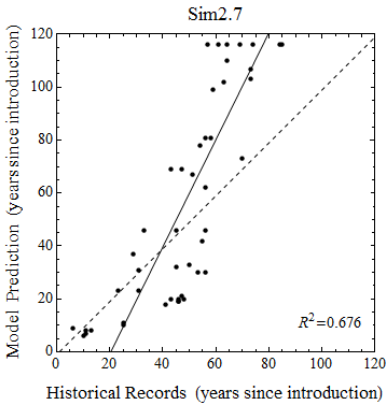


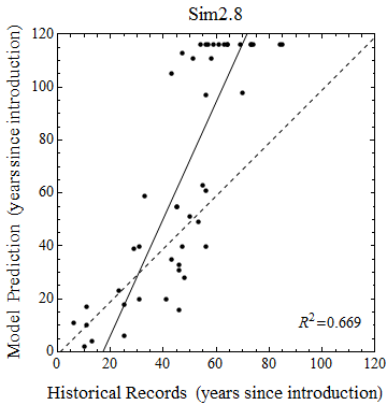


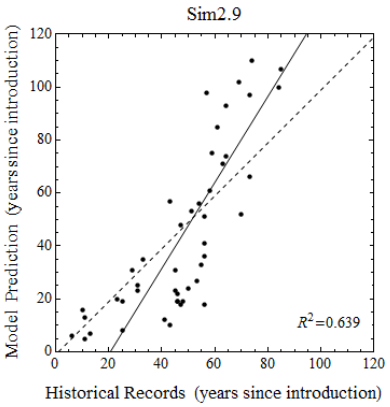


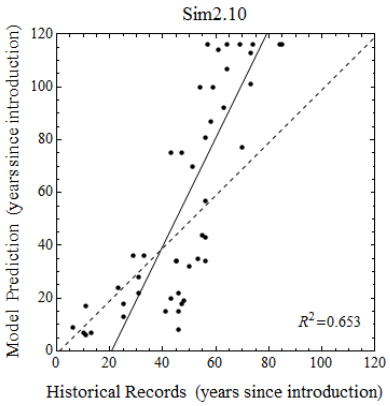


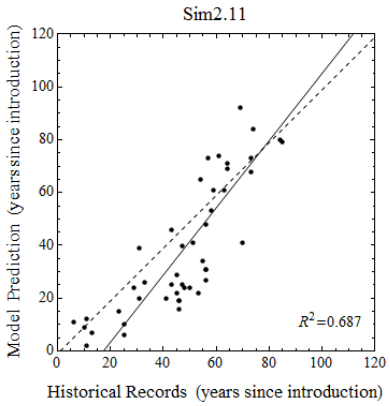


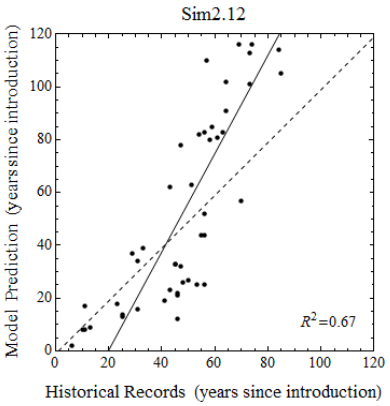


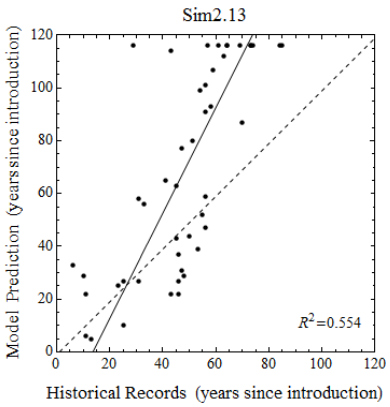


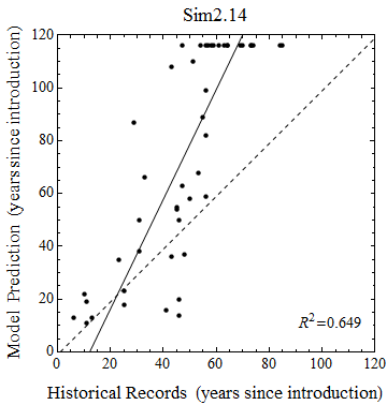


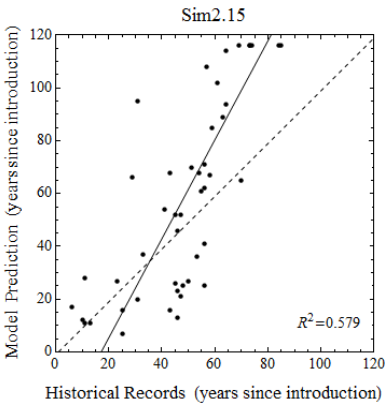


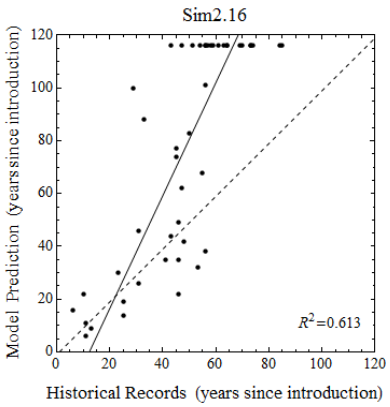


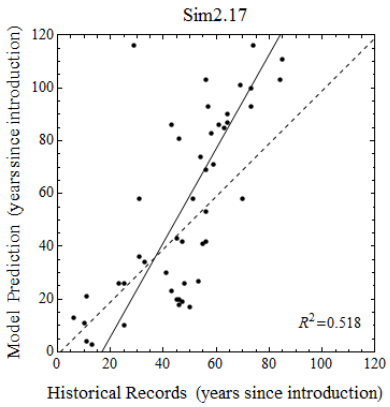


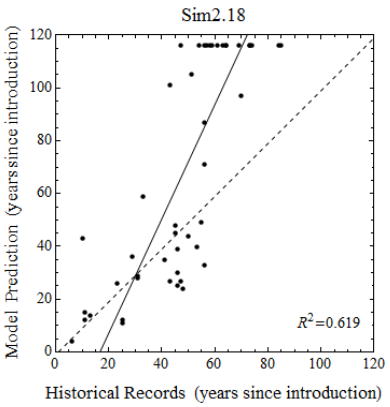


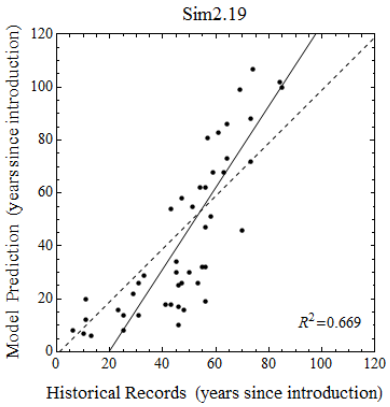


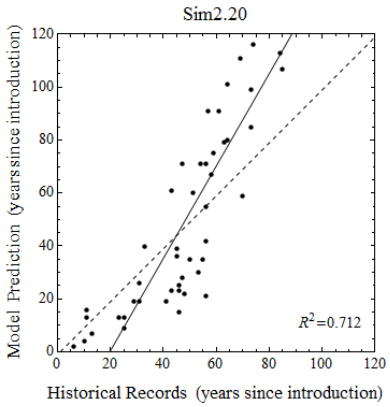


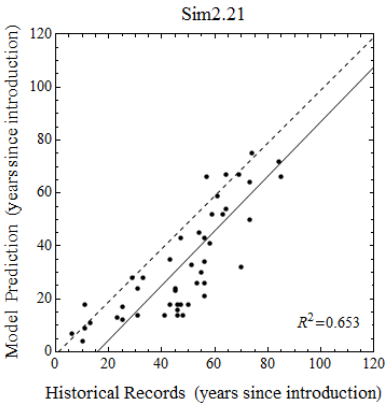


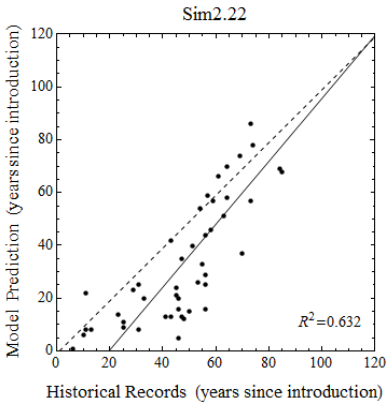


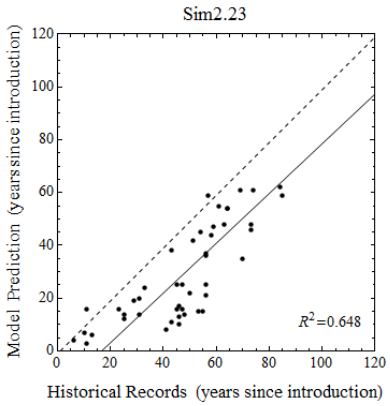


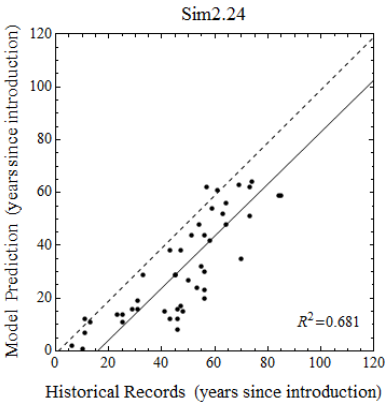


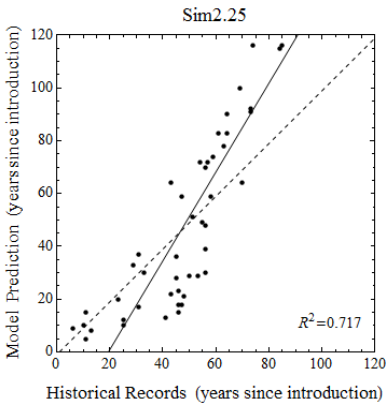


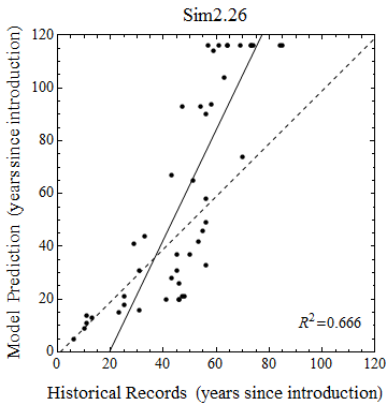


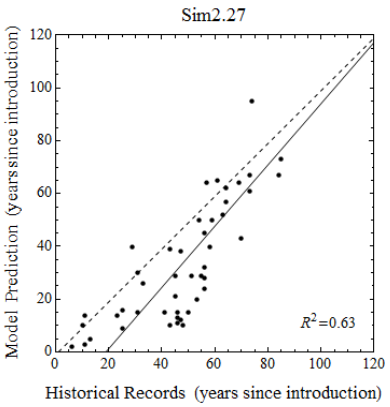


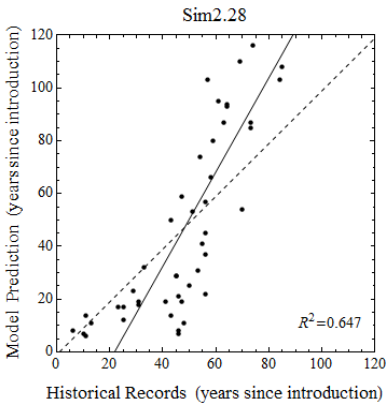


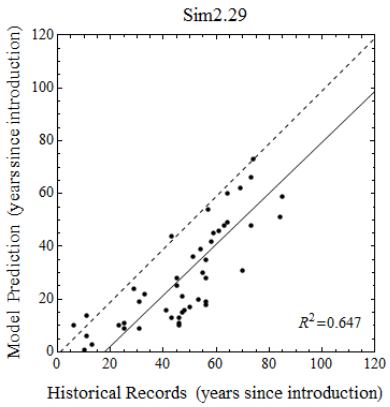


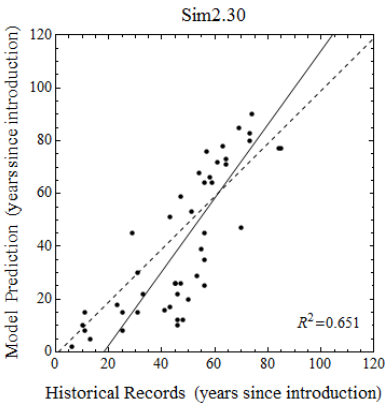


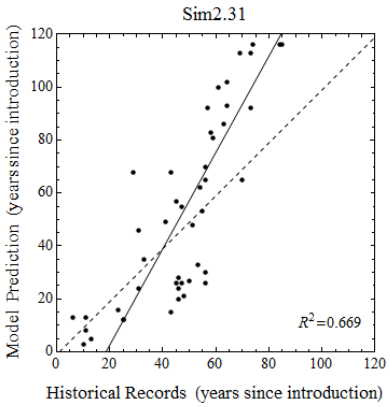


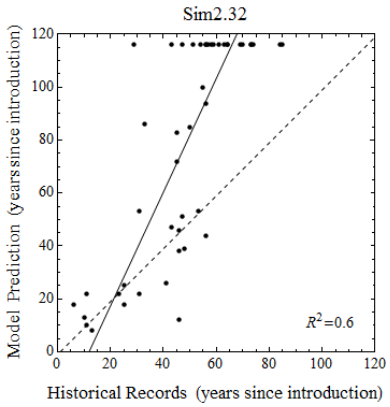


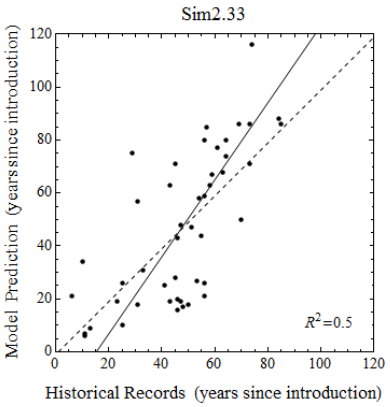


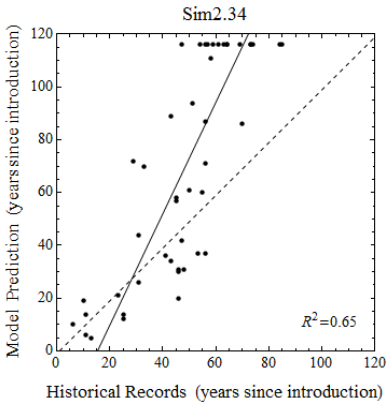


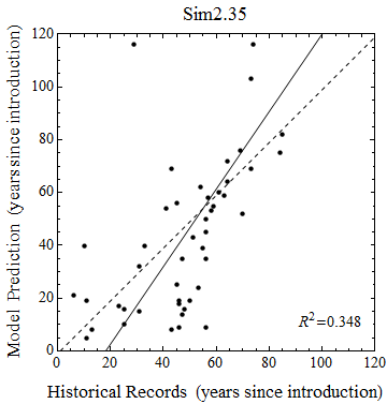


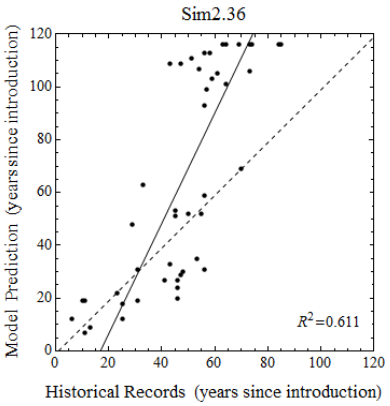


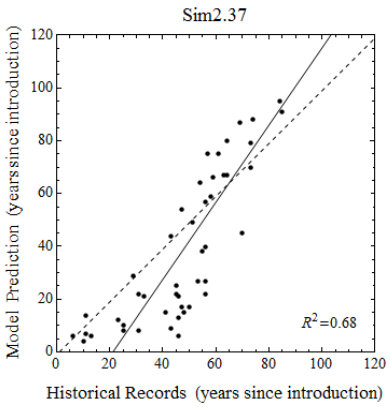


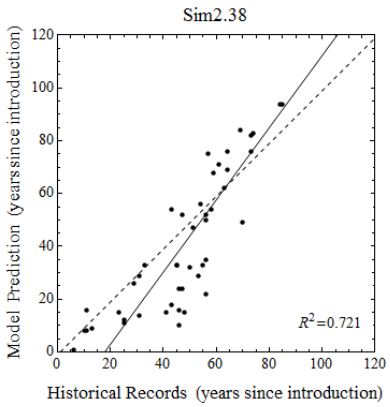


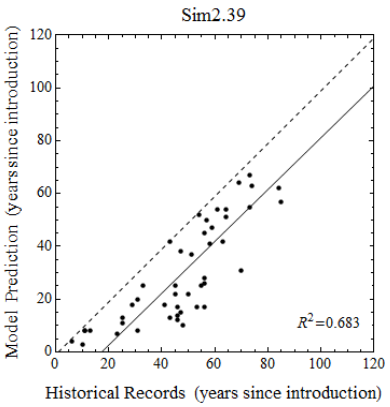


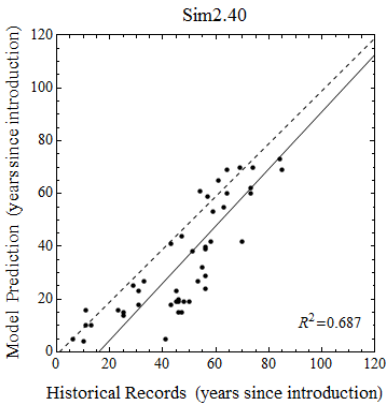


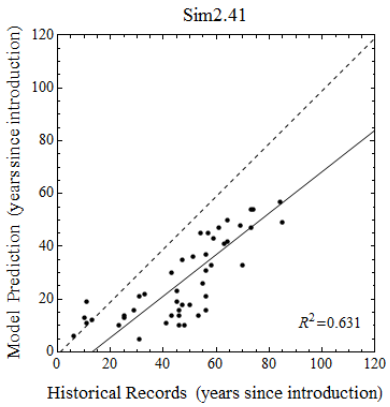


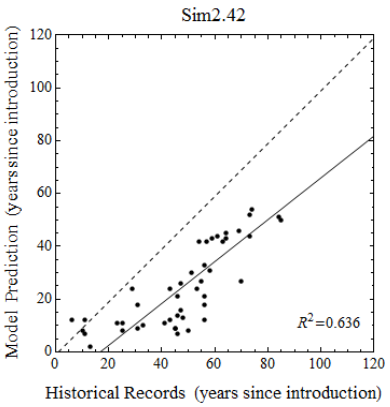


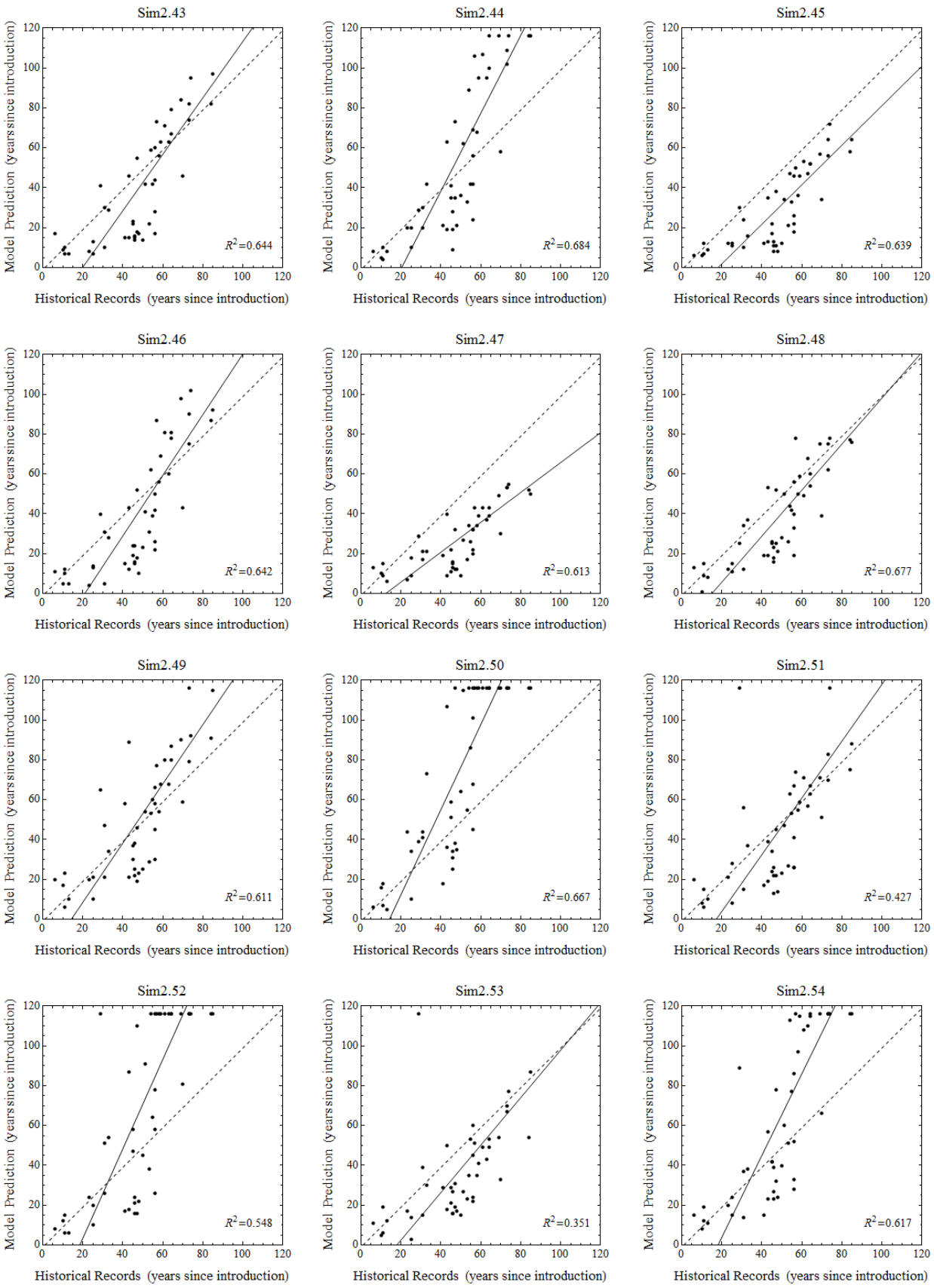












APPENDIX E: Current Records' Evaluation Results

	Evaluation Criteria						Rankings				
Model	Overall Accuracy	Sensitivity	Specificity	Kappa	TSS	AUC	Overall Accuracy	Kappa	TSS	AUC	Sum Rank
Sim1.1	0.796	0.549	0.891	0.465	0.441	0.737	26	34.5	50	68	38
Sim1.2	0.784	0.493	0.896	0.42	0.389	0.708	55	49	59	78	58
Sim1.3	0.804	0.649	0.863	0.511	0.512	0.78	20	19	30	59	18.5
Sim1.4	0.789	0.611	0.858	0.471	0.469	0.76	44	31.5	37	61.5	35
Sim1.5	0.821	0.814	0.824	0.588	0.637	0.861	8	7	4	13.5	4
Sim1.6	0.817	0.866	0.798	0.593	0.664	0.882	9.5	6	1	10	1
Sim1.7	0.787	0.385	0.941	0.381	0.326	0.669	50	63	74	88	72
Sim1.8	0.793	0.355	0.962	0.38	0.316	0.662	35	64	78	91.5	71
Sim1.9	0.813	0.531	0.921	0.492	0.452	0.736	13	25.5	45	69	27
Sim1.10	0.794	0.449	0.926	0.422	0.375	0.696	32	47.5	61.5	80	52
Sim1.11	0.858	0.719	0.911	0.64	0.63	0.828	1	1	7	34.5	6
Sim1.12	0.784	0.461	0.908	0.407	0.369	0.695	55	51	64	81	65
Sim1.13	0.788	0.247	0.996	0.315	0.243	0.621	48.5	83	96	104	95
Sim1.14	0.772	0.192	0.995	0.247	0.187	0.593	66.5	91	103	108	108
Sim1.15	0.792	0.333	0.968	0.367	0.301	0.653	36	70.5	83	95.5	77.5
Sim1.16	0.783	0.235	0.994	0.297	0.228	0.614	58.5	85	97	105	100
Sim1.17	0.795	0.367	0.959	0.389	0.326	0.666	28	58	74	90	64
Sim1.18	0.8	0.315	0.987	0.377	0.302	0.651	23	65	81.5	97	70
Sim1.19	0.813	0.683	0.863	0.54	0.546	0.796	13	12.5	21	54	14
Sim1.20	0.789	0.575	0.871	0.459	0.447	0.743	44	40	47	65	46
Sim1.21	0.784	0.888	0.744	0.54	0.632	0.889	55	12.5	5.5	5	11
Sim1.22	0.789	0.707	0.82	0.5	0.527	0.799	44	22.5	25.5	53	26
Sim1.23	0.638	0.936	0.523	0.337	0.459	0.885	86	79	42	8	50
Sim1.24	0.771	0.894	0.724	0.52	0.618	0.888	68.5	17	10	6	15
Sim1.25	0.817	0.533	0.925	0.5	0.458	0.738	9.5	22.5	43.5	67	25
Sim1.26	0.788	0.395	0.939	0.387	0.334	0.674	48.5	59.5	72	84.5	69
Sim1.27	0.846	0.744	0.885	0.621	0.629	0.837	3	3	8	28	5
Sim1.28	0.786	0.461	0.911	0.41	0.372	0.697	51	50	63	79	61.5
Sim1.29	0.844	0.784	0.868	0.626	0.651	0.852	4	2	2	20	2
Sim1.30	0.789	0.525	0.89	0.441	0.415	0.723	44	44	57	75	51
Sim1.31	0.791	0.335	0.967	0.367	0.302	0.653	37.5	70.5	81.5	95.5	77.5
Sim1.32	0.78	0.218	0.995	0.28	0.214	0.606	61.5	87	100	107	104
Sim1.33	0.795	0.381	0.955	0.397	0.335	0.671	28	56	71	86.5	59.5
Sim1.34	0.791	0.281	0.988	0.34	0.268	0.635	37.5	78	94	102	89
Sim1.35	0.812	0.485	0.938	0.474	0.423	0.717	16	29.5	54	76	36
Sim1.36	0.798	0.341	0.973	0.383	0.314	0.66	25	62	79	93	68
Sim1.37	0.824	0.77	0.845	0.584	0.615	0.838	6.5	8	12	26.5	7
Sim1.38	0.794	0.613	0.863	0.481	0.476	0.76	32	27	36	61.5	28
Sim1.39	0.726	0.942	0.643	0.461	0.585	0.906	75	37	15	1	18.5
Sim1.40	0.778	0.902	0.731	0.533	0.632	0.895	64	15	5.5	3.5	12

Sim1.41	0.584	0.96	0.44	0.278	0.399	0.886	90	88	58	7	61.5
Sim1.42	0.66	0.946	0.55	0.369	0.496	0.901	83	69	32	2	41
Sim1.43	0.815	0.515	0.931	0.492	0.446	0.731	11	25.5	48	72.5	29
Sim1.44	0.794	0.397	0.946	0.4	0.343	0.677	32	53	66	83	56
Sim1.45	0.83	0.796	0.844	0.602	0.639	0.854	5	5	3	18	3
Sim1.46	0.812	0.597	0.894	0.511	0.491	0.762	16	19	34	60	20
Sim1.47	0.803	0.82	0.797	0.557	0.616	0.857	21	9	11	16.5	8
Sim1.48	0.812	0.665	0.868	0.532	0.534	0.79	16	16	23	55	17
Sim1.49	0.789	0.365	0.952	0.376	0.317	0.662	44	66	77	91.5	74
Sim1.50	0.78	0.222	0.995	0.284	0.217	0.608	61.5	86	98	106	102
Sim1.51	0.824	0.515	0.943	0.511	0.458	0.735	6.5	19	43.5	70	24
Sim1.52	0.801	0.329	0.982	0.385	0.311	0.657	22	61	80	94	67
Sim1.53	0.85	0.665	0.921	0.61	0.586	0.804	2	4	14	48.5	9
Sim1.54	0.79	0.317	0.972	0.356	0.289	0.647	39.5	74	85	98	84
Sim2.1	0.771	0.748	0.78	0.48	0.527	0.804	68.5	28	25.5	48.5	34
Sim2.2	0.765	0.705	0.788	0.457	0.493	0.785	71	41	33	58	48
Sim2.3	0.492	0.968	0.309	0.179	0.277	0.838	99	100	92	26.5	90
Sim2.4	0.635	0.93	0.521	0.331	0.451	0.868	87	82	46	12	54
Sim2.5	0.489	0.984	0.299	0.182	0.283	0.861	100	97	87	13.5	85
Sim2.6	0.493	0.968	0.31	0.18	0.278	0.844	97.5	98.5	90.5	23	88
Sim2.7	0.77	0.601	0.835	0.432	0.436	0.741	70	45	52	66	55
Sim2.8	0.779	0.403	0.923	0.371	0.326	0.671	63	67.5	74	86.5	79
Sim2.9	0.709	0.894	0.637	0.422	0.531	0.853	78	47.5	24	19	32
Sim2.10	0.772	0.641	0.822	0.449	0.463	0.757	66.5	42	39	63	49
Sim2.11	0.567	0.958	0.417	0.258	0.375	0.828	91	90	61.5	34.5	73
Sim2.12	0.773	0.816	0.757	0.503	0.572	0.836	65	21	17	29	22
Sim2.13	0.805	0.491	0.925	0.461	0.416	0.713	19	37	56	77	43
Sim2.14	0.789	0.283	0.984	0.336	0.266	0.634	44	80	95	103	92
Sim2.15	0.813	0.679	0.864	0.537	0.543	0.786	13	14	22	56.5	16
Sim2.16	0.783	0.305	0.967	0.334	0.272	0.638	58.5	81	93	101	97
Sim2.17	0.756	0.818	0.732	0.474	0.55	0.819	73	29.5	19	43	31
Sim2.18	0.781	0.389	0.932	0.371	0.321	0.668	60	67.5	76	89	80.5
Sim2.19	0.691	0.922	0.603	0.405	0.525	0.875	79	52	27	11	33
Sim2.20	0.761	0.856	0.724	0.493	0.58	0.851	72	24	16	21	23
Sim2.21	0.427	0.99	0.211	0.123	0.201	0.829	102	102	101	31.5	98
Sim2.22	0.437	0.992	0.224	0.133	0.216	0.847	101	101	99	22	93
Sim2.23	0.39	0.988	0.161	0.089	0.149	0.825	106	106	106	40	105
Sim2.24	0.393	0.992	0.163	0.093	0.155	0.826	105	105	105	38.5	103
Sim2.25	0.753	0.824	0.725	0.471	0.549	0.827	74	31.5	20	37	30
Sim2.26	0.785	0.559	0.872	0.446	0.431	0.73	52	43	53	74	53
Sim2.27	0.542	0.96	0.381	0.23	0.341	0.82	92.5	92	67	42	82
Sim2.28	0.718	0.854	0.665	0.424	0.519	0.829	77	46	28	31.5	39.5
Sim2.29	0.501	0.966	0.322	0.188	0.288	0.804	96	96	86	48.5	94
Sim2.30	0.542	0.954	0.384	0.228	0.338	0.83	92.5	93	69	30	76
Sim2.31	0.809	0.732	0.839	0.545	0.571	0.804	18	11	18	48.5	13
Sim2.32	0.79	0.303	0.977	0.348	0.28	0.641	39.5	77	89	100	86
Sim2.33	0.669	0.868	0.592	0.357	0.46	0.806	82	73	41	45.5	59.5

Sim2.34	0.794	0.391	0.949	0.399	0.34	0.674	32	54.5	68	84.5	57
Sim2.35	0.619	0.912	0.507	0.306	0.418	0.802	88	84	55	52	75
Sim2.36	0.794	0.555	0.885	0.461	0.44	0.731	32	37	51	72.5	45
Sim2.37	0.655	0.946	0.543	0.362	0.489	0.884	84.5	72	35	9	47
Sim2.38	0.724	0.95	0.637	0.46	0.587	0.895	76	39	13	3.5	21
Sim2.39	0.411	0.992	0.188	0.109	0.18	0.828	104	104	104	34.5	101
Sim2.40	0.418	0.99	0.198	0.115	0.188	0.839	103	103	102	25	96
Sim2.41	0.385	0.99	0.152	0.085	0.142	0.824	107	107	107	41	106
Sim2.42	0.384	0.99	0.151	0.084	0.141	0.826	108	108	108	38.5	107
Sim2.43	0.676	0.934	0.577	0.387	0.51	0.858	81	59.5	31	15	42
Sim2.44	0.784	0.621	0.847	0.465	0.468	0.756	55	34.5	38	64	44
Sim2.45	0.503	0.964	0.326	0.19	0.29	0.806	95	95	84	45.5	91
Sim2.46	0.677	0.942	0.575	0.391	0.517	0.857	80	57	29	16.5	39.5
Sim2.47	0.493	0.966	0.312	0.18	0.278	0.803	97.5	98.5	90.5	51	99
Sim2.48	0.536	0.966	0.371	0.226	0.337	0.828	94	94	70	34.5	80.5
Sim2.49	0.795	0.844	0.777	0.549	0.62	0.841	28	10	9	24	10
Sim2.50	0.789	0.307	0.975	0.349	0.281	0.642	44	76	88	99	87
Sim2.51	0.655	0.902	0.56	0.35	0.462	0.813	84.5	75	40	44	63
Sim2.52	0.784	0.445	0.914	0.399	0.359	0.688	55	54.5	65	82	66
Sim2.53	0.59	0.914	0.466	0.271	0.38	0.786	89	89	60	56.5	83
Sim2.54	0.799	0.545	0.897	0.469	0.442	0.732	24	33	49	71	37

APPENDIX F: Selected Evaluation Results for Historical and Current Records' Evaluations and their Respective Rankings

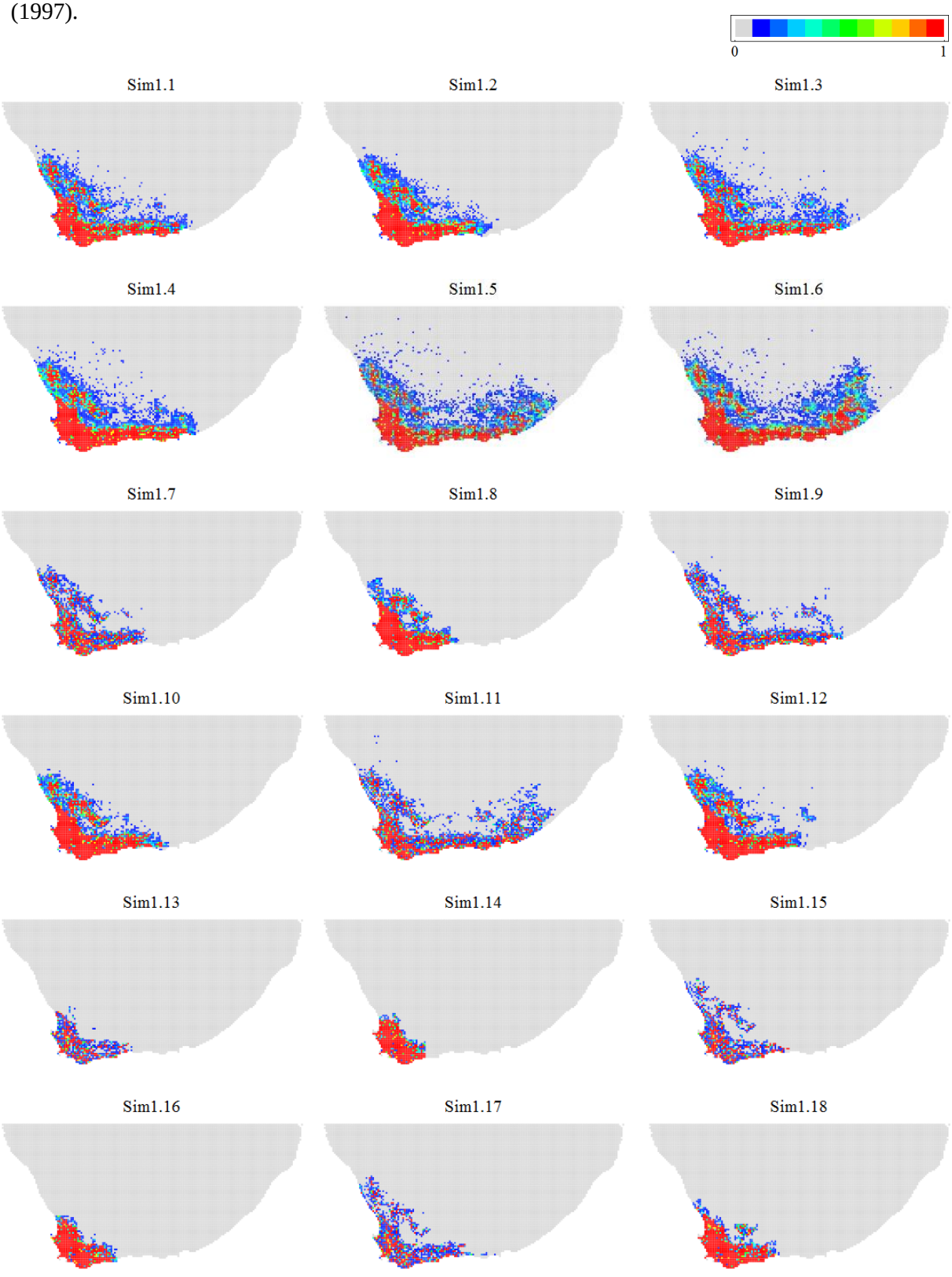
	Historical Records' Evaluations			Current Records' Evaluations				Rankings		
Model	RMA Slope	Absolute Difference of Slope to 1	R^2	Overall Accuracy	Kappa	TSS	AUC	Historical Records' Evaluations	Current Records' Evaluations	Combined Evaluations (Slope, R^2, Kappa & TSS)
Sim1.1	2.018	1.018	0.665	0.796	0.465	0.441	0.737	61	38	45
Sim1.2	2.121	1.121	0.668	0.784	0.42	0.389	0.708	69	58	56.5
Sim1.3	1.901	0.901	0.678	0.804	0.511	0.512	0.78	34	18.5	22
Sim1.4	1.918	0.918	0.667	0.789	0.471	0.469	0.76	49	35	35
Sim1.5	1.650	0.650	0.656	0.821	0.588	0.637	0.861	39	4	12
Sim1.6	1.637	0.637	0.645	0.817	0.593	0.664	0.882	54	1	13
Sim1.7	2.131	1.131	0.689	0.787	0.381	0.326	0.669	54	72	61
Sim1.8	2.169	1.169	0.698	0.793	0.38	0.316	0.662	57	71	67.5
Sim1.9	1.871	0.871	0.651	0.813	0.492	0.452	0.736	61	27	37
Sim1.10	2.129	1.129	0.658	0.794	0.422	0.375	0.696	76.5	52	63
Sim1.11	1.680	0.680	0.691	0.858	0.64	0.63	0.828	15	6	2
Sim1.12	2.211	1.211	0.608	0.784	0.407	0.369	0.695	104.5	65	91
Sim1.13	2.058	1.058	0.57	0.788	0.315	0.243	0.621	92.5	95	104.5
Sim1.14	2.089	1.089	0.602	0.772	0.247	0.187	0.593	92.5	108	108
Sim1.15	2.190	1.190	0.604	0.792	0.367	0.301	0.653	104.5	77.5	101.5
Sim1.16	2.162	1.162	0.607	0.783	0.297	0.228	0.614	102	100	107
Sim1.17	2.023	1.023	0.514	0.795	0.389	0.326	0.666	97.5	64	92
Sim1.18	2.176	1.176	0.566	0.8	0.377	0.302	0.651	107	70	101.5
Sim1.19	1.862	0.862	0.697	0.813	0.54	0.546	0.796	17	14	10
Sim1.20	1.984	0.984	0.716	0.789	0.459	0.447	0.743	23	46	30
Sim1.21	1.465	0.465	0.624	0.784	0.54	0.632	0.889	61	11	19
Sim1.22	1.843	0.843	0.648	0.789	0.5	0.527	0.799	58	26	29
Sim1.23	1.285	0.285	0.633	0.638	0.337	0.459	0.885	44	50	52
Sim1.24	1.562	0.562	0.657	0.771	0.52	0.618	0.888	34	15	14
Sim1.25	1.881	0.881	0.704	0.817	0.5	0.458	0.738	20	25	21
Sim1.26	2.241	1.241	0.624	0.788	0.387	0.334	0.674	101	69	94
Sim1.27	1.554	0.554	0.644	0.846	0.621	0.629	0.837	50	5	15
Sim1.28	2.180	1.180	0.65	0.786	0.41	0.372	0.697	85.5	61.5	75
Sim1.29	1.398	0.398	0.642	0.844	0.626	0.651	0.852	40.5	2	6.5
Sim1.30	2.004	1.004	0.692	0.789	0.441	0.415	0.723	30	51	38
Sim1.31	2.103	1.103	0.664	0.791	0.367	0.302	0.653	71.5	77.5	79.5
Sim1.32	2.063	1.063	0.624	0.78	0.28	0.214	0.606	87	104	100
Sim1.33	2.077	1.077	0.582	0.795	0.397	0.335	0.671	95	59.5	88
Sim1.34	2.107	1.107	0.654	0.791	0.34	0.268	0.635	76.5	89	90
Sim1.35	1.864	0.864	0.639	0.812	0.474	0.423	0.717	71.5	36	50
Sim1.36	2.173	1.173	0.588	0.798	0.383	0.314	0.66	106	68	98
Sim1.37	1.703	0.703	0.699	0.824	0.584	0.615	0.838	12.5	7	3
Sim1.38	1.986	0.986	0.667	0.794	0.481	0.476	0.76	54	28	32

Sim1.39	1.125	0.125	0.671	0.726	0.461	0.585	0.906	8	18.5	5
Sim1.40	1.404	0.404	0.651	0.778	0.533	0.632	0.895	25	12	8
Sim1.41	1.051	0.051	0.699	0.584	0.278	0.399	0.886	1	61.5	31
Sim1.42	1.243	0.243	0.595	0.66	0.369	0.496	0.901	61	41	51
Sim1.43	1.900	0.900	0.688	0.815	0.492	0.446	0.731	26.5	29	26
Sim1.44	2.274	1.274	0.658	0.794	0.4	0.343	0.677	84	56	76.5
Sim1.45	1.428	0.428	0.677	0.83	0.602	0.639	0.854	12.5	3	1
Sim1.46	1.868	0.868	0.679	0.812	0.511	0.491	0.762	28	20	20
Sim1.47	1.323	0.323	0.65	0.803	0.557	0.616	0.857	26.5	8	9
Sim1.48	1.870	0.870	0.677	0.812	0.532	0.534	0.79	29	17	17
Sim1.49	2.092	1.092	0.603	0.789	0.376	0.317	0.662	95	74	95
Sim1.50	2.144	1.144	0.61	0.78	0.284	0.217	0.608	100	102	106
Sim1.51	1.738	0.738	0.616	0.824	0.511	0.458	0.735	74	24	43
Sim1.52	2.058	1.058	0.66	0.801	0.385	0.311	0.657	66	67	72
Sim1.53	1.712	0.712	0.542	0.85	0.61	0.586	0.804	83	9	33.5
Sim1.54	2.222	1.222	0.631	0.79	0.356	0.289	0.647	99	84	99
Sim2.1	1.929	0.929	0.672	0.771	0.48	0.527	0.804	44	34	24
Sim2.2	1.962	0.962	0.672	0.765	0.457	0.493	0.785	46	48	33.5
Sim2.3	1.471	0.471	0.644	0.492	0.179	0.277	0.838	47	90	84
Sim2.4	1.552	0.552	0.642	0.635	0.331	0.451	0.868	56	54	58
Sim2.5	1.248	0.248	0.66	0.489	0.182	0.283	0.861	16	85	62
Sim2.6	1.524	0.524	0.648	0.493	0.18	0.278	0.844	44	88	82
Sim2.7	2.024	1.024	0.676	0.77	0.432	0.436	0.741	48	55	46
Sim2.8	2.207	1.207	0.669	0.779	0.371	0.326	0.671	76.5	79	78
Sim2.9	1.625	0.625	0.639	0.709	0.422	0.531	0.853	64	32	39
Sim2.10	2.054	1.054	0.653	0.772	0.449	0.463	0.757	70	49	49
Sim2.11	1.273	0.273	0.687	0.567	0.258	0.375	0.828	7	73	41
Sim2.12	1.862	0.862	0.67	0.773	0.503	0.572	0.836	32	22	18
Sim2.13	1.998	0.998	0.554	0.805	0.461	0.416	0.713	90.5	43	73
Sim2.14	2.079	1.079	0.649	0.789	0.336	0.266	0.634	80	92	93
Sim2.15	1.876	0.876	0.579	0.813	0.537	0.543	0.786	88.5	16	44
Sim2.16	2.140	1.140	0.613	0.783	0.334	0.272	0.638	95	97	103
Sim2.17	1.782	0.782	0.518	0.756	0.474	0.55	0.819	85.5	31	48
Sim2.18	2.165	1.165	0.619	0.781	0.371	0.321	0.668	97.5	80.5	96
Sim2.19	1.542	0.542	0.669	0.691	0.405	0.525	0.875	21	33	25
Sim2.20	1.745	0.745	0.712	0.761	0.493	0.58	0.851	14	23	6.5
Sim2.21	1.032	0.032	0.653	0.427	0.123	0.201	0.829	11	98	70
Sim2.22	1.192	0.192	0.632	0.437	0.133	0.216	0.847	36.5	93	83
Sim2.23	0.943	0.057	0.648	0.39	0.089	0.149	0.825	18.5	105	79.5
Sim2.24	0.985	0.015	0.681	0.393	0.093	0.155	0.826	2.5	103	56.5
Sim2.25	1.683	0.683	0.717	0.753	0.471	0.549	0.827	9	30	11
Sim2.26	2.085	1.085	0.666	0.785	0.446	0.431	0.73	66	53	53
Sim2.27	1.160	0.160	0.63	0.542	0.23	0.341	0.82	38	82	64
Sim2.28	1.786	0.786	0.647	0.718	0.424	0.519	0.829	61	39.5	40
Sim2.29	0.967	0.033	0.647	0.501	0.188	0.288	0.804	18.5	94	65
Sim2.30	1.394	0.394	0.651	0.542	0.228	0.338	0.83	24	76	59
Sim2.31	1.820	0.820	0.669	0.809	0.545	0.571	0.804	34	13	16

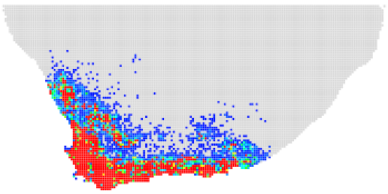
Sim2.32	2.149	1.149	0.6	0.79	0.348	0.28	0.641	103	86	104.5
Sim2.33	1.456	0.456	0.5	0.669	0.357	0.46	0.806	79	59.5	67.5
Sim2.34	2.112	1.112	0.65	0.794	0.399	0.34	0.674	82	57	74
Sim2.35	1.469	0.469	0.348	0.619	0.306	0.418	0.802	81	75	81
Sim2.36	2.091	1.091	0.611	0.794	0.461	0.44	0.731	90.5	45	71
Sim2.37	1.462	0.462	0.68	0.655	0.362	0.489	0.884	10	47	28
Sim2.38	1.367	0.367	0.721	0.724	0.46	0.587	0.895	5	21	4
Sim2.39	0.982	0.018	0.683	0.411	0.109	0.18	0.828	2.5	101	55
Sim2.40	1.082	0.082	0.687	0.418	0.115	0.188	0.839	4	96	54
Sim2.41	0.786	0.214	0.631	0.385	0.085	0.142	0.824	42	106	89
Sim2.42	0.793	0.207	0.636	0.384	0.084	0.141	0.826	36.5	107	87
Sim2.43	1.415	0.415	0.644	0.676	0.387	0.51	0.858	40.5	42	36
Sim2.44	1.952	0.952	0.684	0.784	0.465	0.468	0.756	31	44	27
Sim2.45	0.991	0.009	0.639	0.503	0.19	0.29	0.806	22	91	67.5
Sim2.46	1.524	0.524	0.642	0.677	0.391	0.517	0.857	51.5	39.5	42
Sim2.47	0.751	0.249	0.613	0.493	0.18	0.278	0.803	51.5	99	85.5
Sim2.48	1.156	0.156	0.677	0.536	0.226	0.337	0.828	6	80.5	47
Sim2.49	1.489	0.489	0.611	0.795	0.549	0.62	0.841	66	10	23
Sim2.50	2.138	1.138	0.667	0.789	0.349	0.281	0.642	73	87	85.5
Sim2.51	1.426	0.426	0.427	0.655	0.35	0.462	0.813	76.5	63	67.5
Sim2.52	2.248	1.248	0.548	0.784	0.399	0.359	0.688	108	66	97
Sim2.53	1.192	0.192	0.351	0.59	0.271	0.38	0.786	68	83	76.5
Sim2.54	2.054	1.054	0.617	0.799	0.469	0.442	0.732	88.5	37	60

APPENDIX G: Relative Density Plots

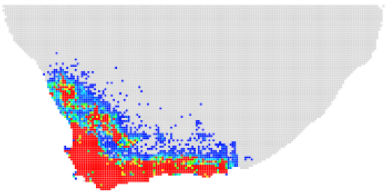
The European Starling's relative density for all 108 model simulations after 100 simulated time steps (1997).



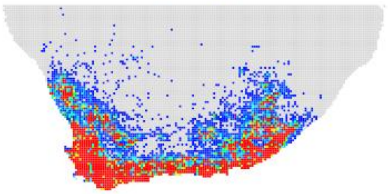
Sim1.19



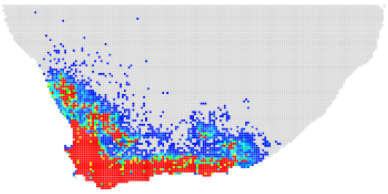
Sim1.20



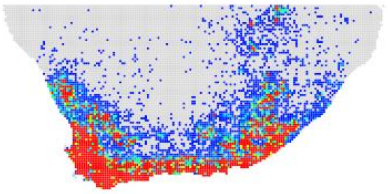
Sim1.21



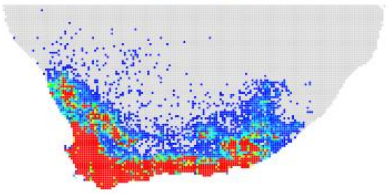
Sim1.22



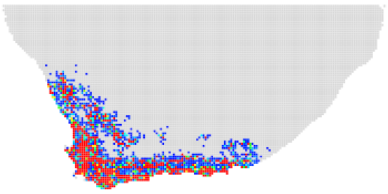
Sim1.23



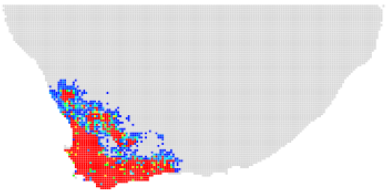
Sim1.24



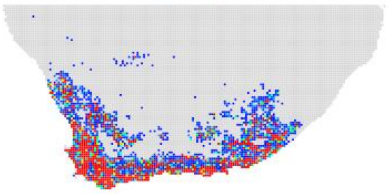
Sim1.25



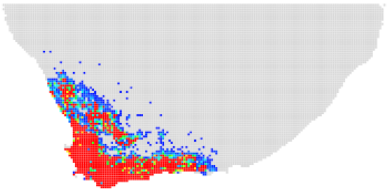
Sim1.26



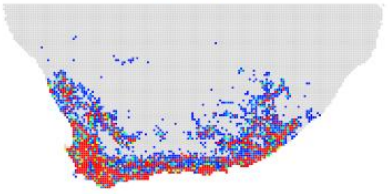
Sim1.27



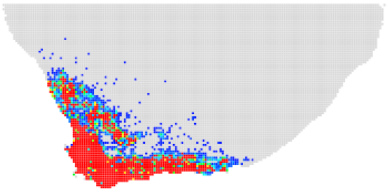
Sim1.28



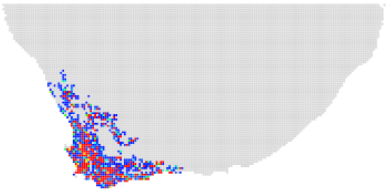
Sim1.29



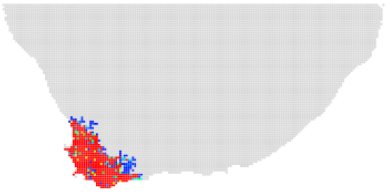
Sim1.30



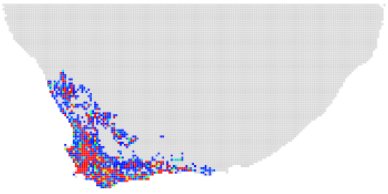
Sim1.31



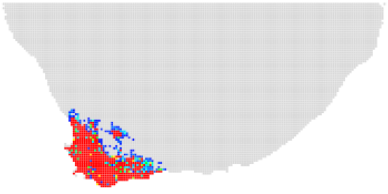
Sim1.32



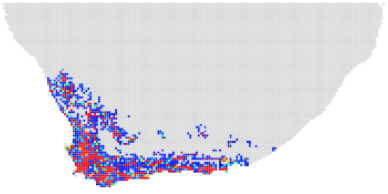
Sim1.33



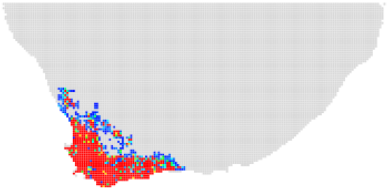
Sim1.34



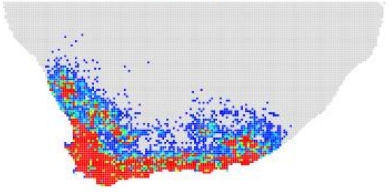
Sim1.35



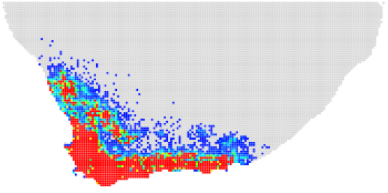
Sim1.36



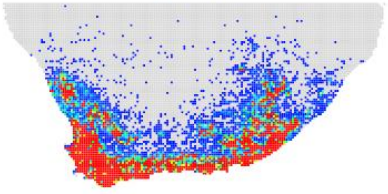
Sim1.37



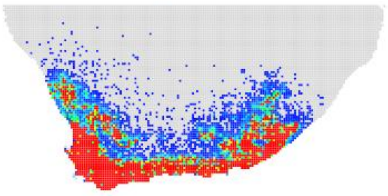
Sim1.38



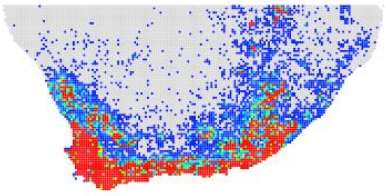
Sim1.39



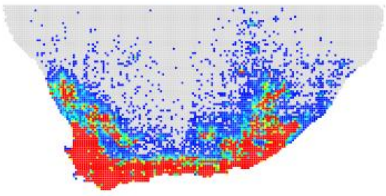
Sim1.40



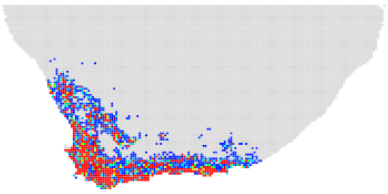
Sim1.41



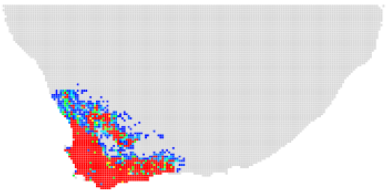
Sim1.42



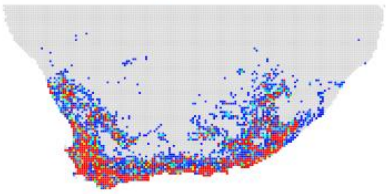
Sim1.43



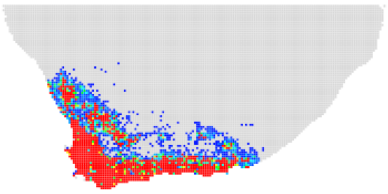
Sim1.44



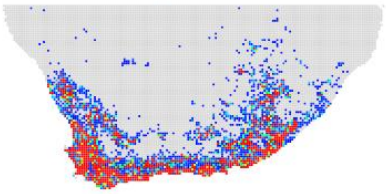
Sim1.45



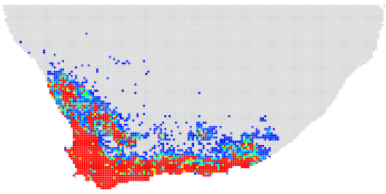
Sim1.46



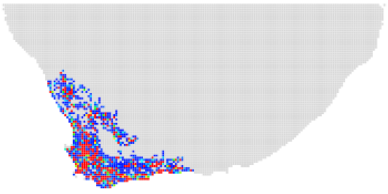
Sim1.47



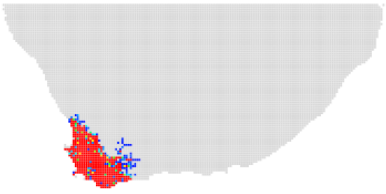
Sim1.48



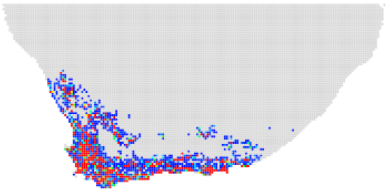
Sim1.49



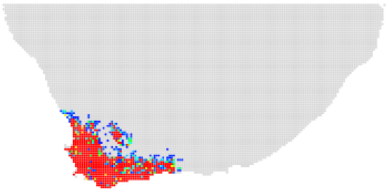
Sim1.50



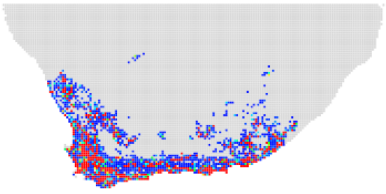
Sim1.51



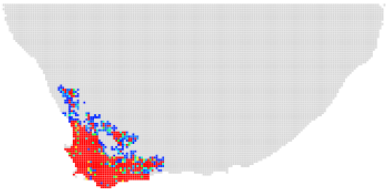
Sim1.52



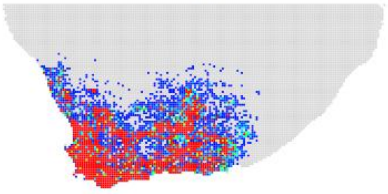
Sim1.53



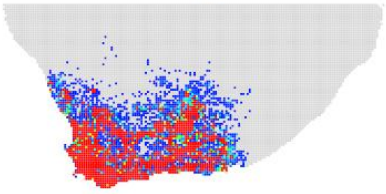
Sim1.54



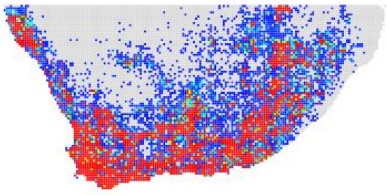
Sim2.1



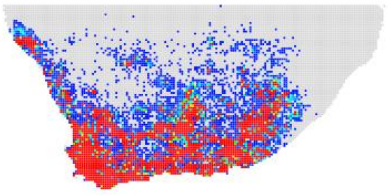
Sim2.2



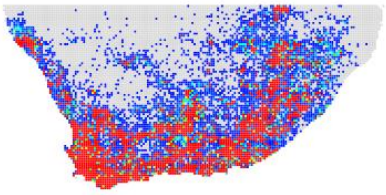
Sim2.3



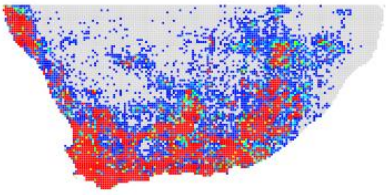
Sim2.4



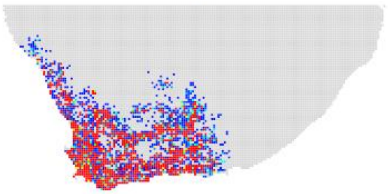
Sim2.5



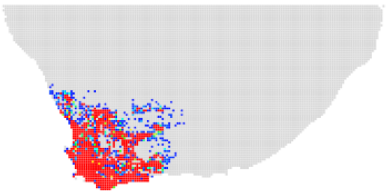
Sim2.6



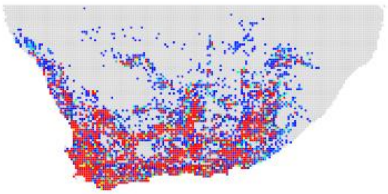
Sim2.7



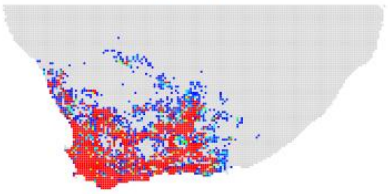
Sim2.8



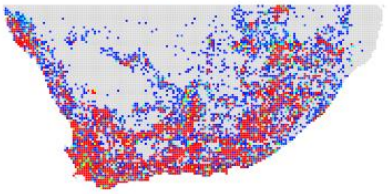
Sim2.9



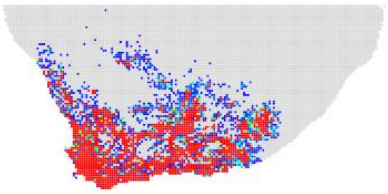
Sim2.10



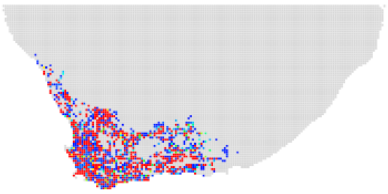
Sim2.11



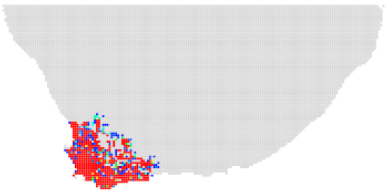
Sim2.12



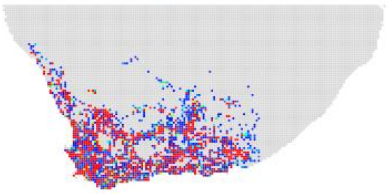
Sim2.13



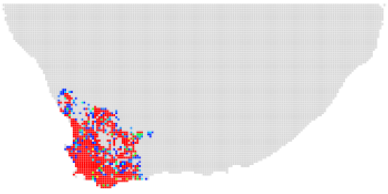
Sim2.14



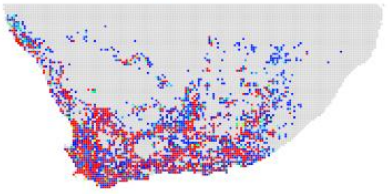
Sim2.15



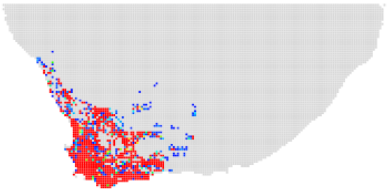
Sim2.16



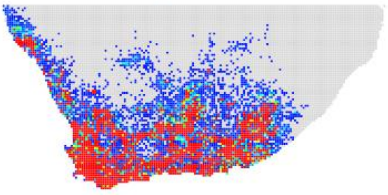
Sim2.17



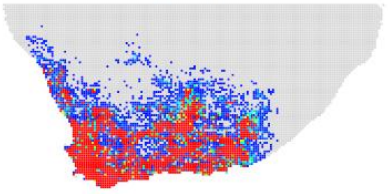
Sim2.18



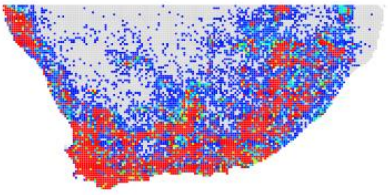
Sim2.19



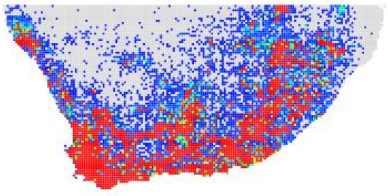
Sim2.20



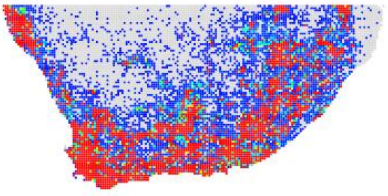
Sim2.21



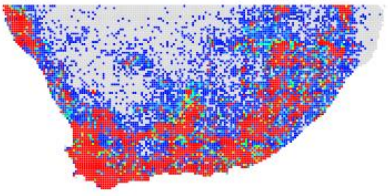
Sim2.22



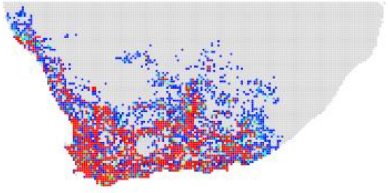
Sim2.23



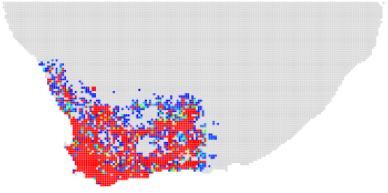
Sim2.24



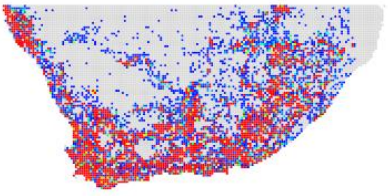
Sim2.25



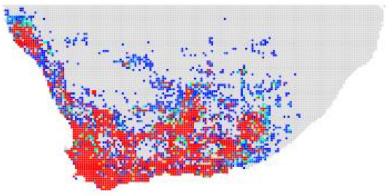
Sim2.26



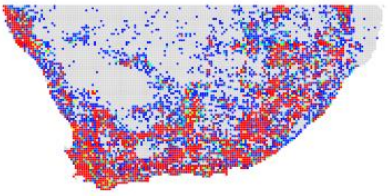
Sim2.27



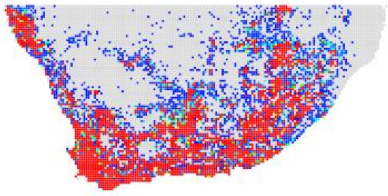
Sim2.28



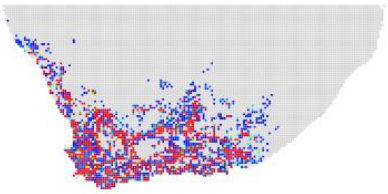
Sim2.29



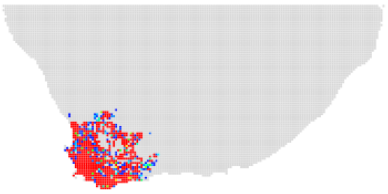
Sim2.30



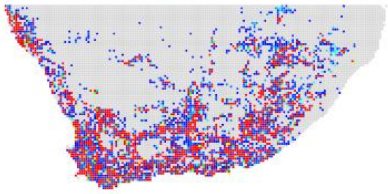
Sim2.31



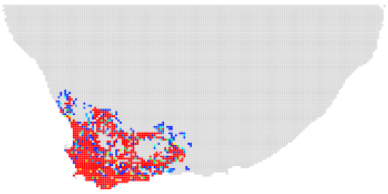
Sim2.32



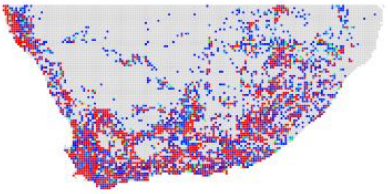
Sim2.33



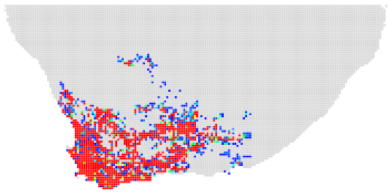
Sim2.34



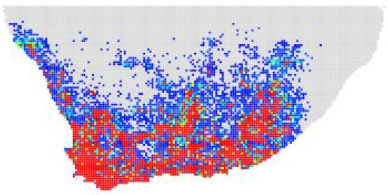
Sim2.35



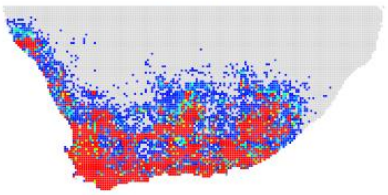
Sim2.36



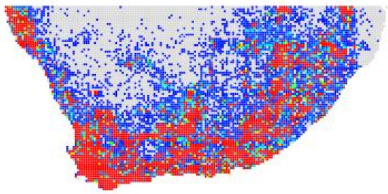
Sim2.37



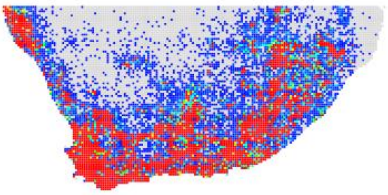
Sim2.38



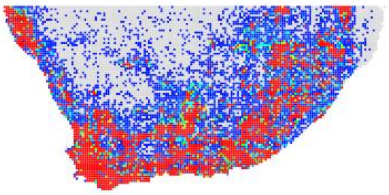
Sim2.39



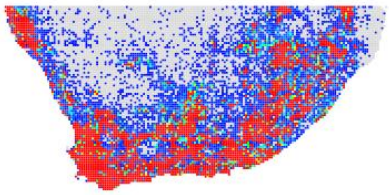
Sim2.40



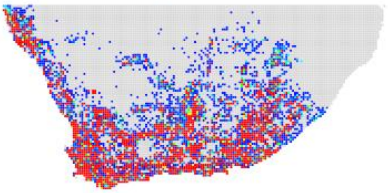
Sim2.41



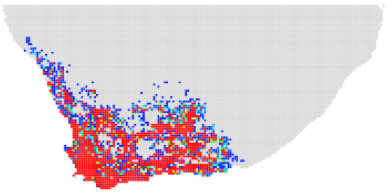
Sim2.42



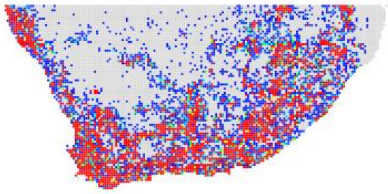
Sim2.43



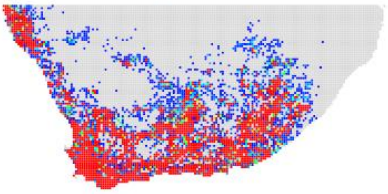
Sim2.44



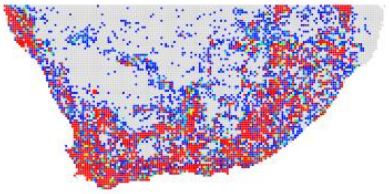
Sim2.45



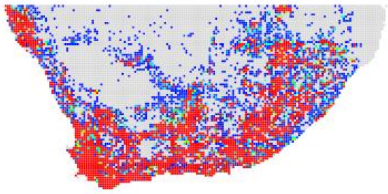
Sim2.46

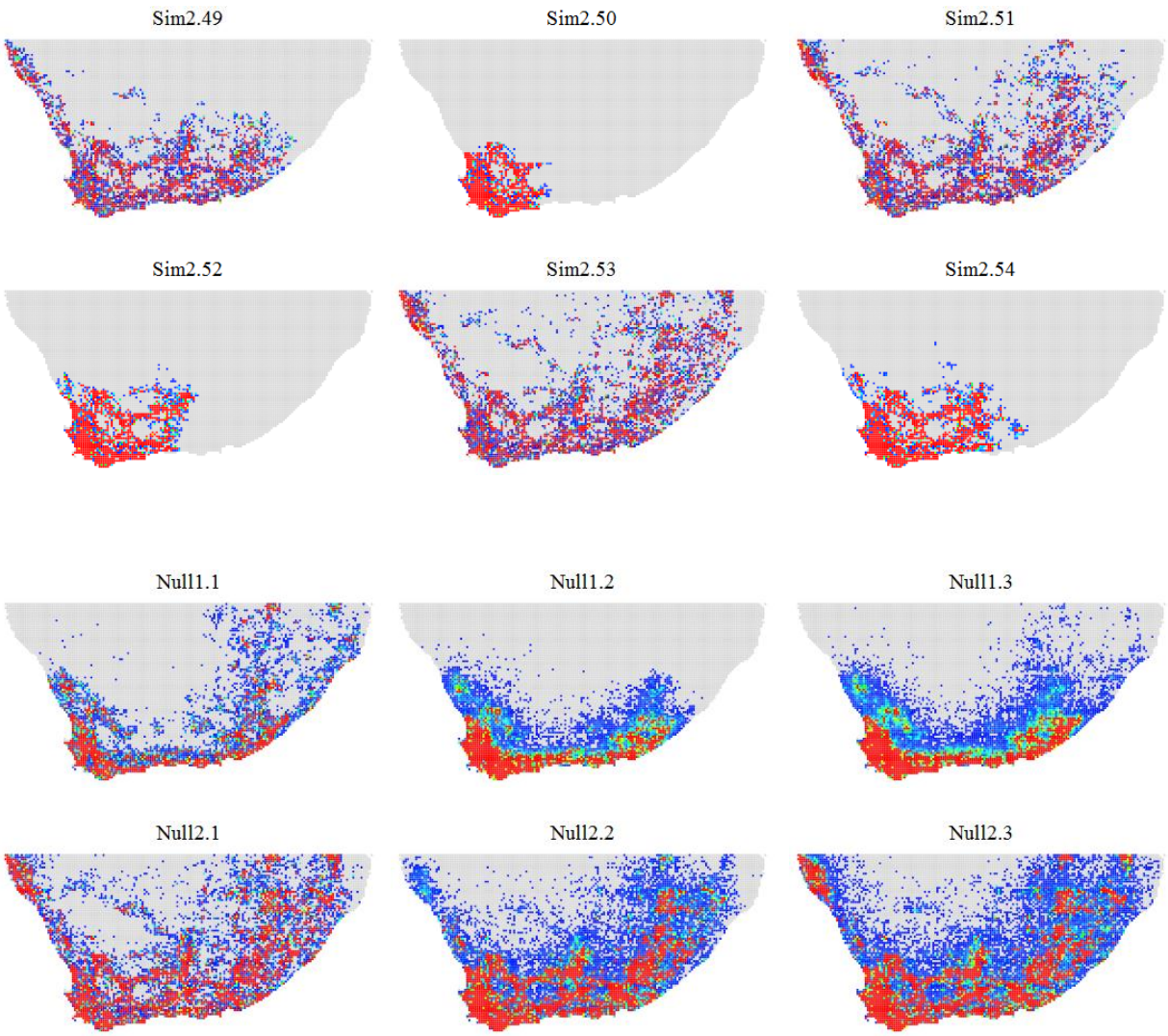


Sim2.47



Sim2.48





APPENDIX H: Parameter Sensitivity Analysis

a.) ANOVA results

For testing whether different parameter levels will result in significantly different model performance, we calculated F-tests for one-way ANOVA at the 5% significance level ($\alpha = 0.05$). The null hypothesis (H_0) states that the sample means are equal and if $F_{obs} > F_{crit}$ (p-value < 0.05) then we can conclude with 95% confidence that the null hypothesis is false and therefore the variance between the groups is not due to random chance. For each parameter we compared two levels among each other (for example 200 m elevation access vs. 300 m elevation access) and did this for both habitat suitability maps combined (SDM1 & SDM2, 108 models) as well as for SDM1 and SDM2 separately (54 models each). We therefore had two critical F-values: $F_{(0.05,1,106)} = 3.94$ and $F_{(0.05,1,52)} = 4.03$.

Table H1: ANOVA results for comparing parameter categories using all 108 model results according to the various evaluation criteria. F refers to the F-ratio and p refers to the p-value. Significant differences in sample means are highlighted in red. $F_{crit} = 3.94$.

			Parameters										
			Habitat Suitability	Elevation Access (m)			Number of Dispersal Choices			Choosing Location	Maximum Dispersal Distance (km)		
			SDM1 vs. SDM2	100 vs. 200	200 vs. 300	100 vs. 300	2 vs. 5	5 vs. 10	2 vs. 10	HS vs. HS DIS	100 vs. 200	200 vs. 300	100 vs. 300
Evaluation Criteria	Slope	F	**14.476**	**6.970**	1.504	**15.783**	**7.157**	**8.527**	**39.460**	**18.321**	**10.710**	1.228	**21.106**
		p	**0.000**	**0.010**	0.224	**0.000**	**0.009**	**0.005**	**0.000**	**0.000**	**0.002**	0.272	**0.000**
	R^2	F	1.707	0.054	0.097	0.018	0.114	**33.239**	**34.563**	3.448	**6.947**	0.740	**8.145**
		p	0.194	0.817	0.757	0.893	0.737	**0.000**	**0.000**	0.066	**0.010**	0.393	**0.006**
	Overall Accuracy	F	**44.097**	2.193	0.185	3.726	**6.300**	**4.049**	**18.936**	1.319	**8.238**	1.034	**14.346**
		p	**0.000**	0.143	0.668	0.058	**0.014**	**0.048**	**0.000**	0.253	**0.005**	0.313	**0.000**
	Kappa	F	**24.611**	0.371	0.006	0.454	2.541	1.184	0.824	0.668	2.541	1.184	0.824
		p	**0.000**	0.544	0.937	0.503	0.115	0.280	0.367	0.416	0.115	0.280	0.367
	TSS	F	3.690	0.031	0.094	0.242	0.212	**9.330**	3.751	**4.874**	0.017	0.029	0.091
		p	0.0574	0.862	0.760	0.624	0.647	**0.003**	0.057	**0.029**	0.896	0.865	0.763
	AUC	F	**8.897**	2.191	0.891	5.707	**18.377**	**23.007**	**86.557**	**11.152**	**4.403**	1.133	**10.434**
		p	**0.004**	0.143	0.348	0.020	**0.000**	**0.000**	**0.000**	**0.001**	**0.039**	0.291	**0.002**

Table H2: ANOVA results for comparing parameter categories using the 54 models that incorporated SDM1's results according to the various evaluation criteria. F refers to the F-ratio and p refers to the p-value. Significant differences in sample means are highlighted in red. F_{crit} = 4.03.

			Parameters									
			Elevation Access (m)			Number of Dispersal Choices			Choosing Location	Maximum Dispersal Distance (km)		
			100 vs. 200	200 vs. 300	100 vs. 300	2 vs. 5	5 vs. 10	2 vs. 10	HS vs. HS DIS	100 vs. 200	200 vs. 300	100 vs. 300
Evaluation Criteria	Slope	F	2.801	1.489	**7.533**	**5.211**	**4.170**	**22.886**	**9.886**	**6.628**	0.883	**11.817**
		p	0.103	0.231	**0.010**	**0.029**	**0.049**	**0.000**	**0.003**	**0.015**	0.354	**0.002**
	R^2	F	0.860	0.001	0.722	0.068	**29.283**	**25.389**	0.001	1.079	1.404	3.352
		p	0.360	0.970	0.401	0.796	**0.000**	**0.000**	0.980	0.306	0.244	0.076
	Overall Accuracy	F	0.553	0.220	1.116	**6.951**	3.625	3.631	0.422	0.123	0.864	0.690
		p	0.462	0.642	0.298	**0.013**	0.065	0.065	0.519	0.728	0.359	0.412
	Kappa	F	0.352	0.110	0.797	0.081	**14.646**	**13.251**	**6.784**	0.081	**14.646**	**13.251**
		p	0.557	0.743	0.378	0.778	**0.001**	**0.001**	**0.012**	0.778	**0.001**	**0.001**
	TSS	F	0.876	0.398	2.446	3.247	**17.820**	**49.811**	**6.063**	**4.383**	0.399	**7.699**
		p	0.356	0.533	0.127	0.080	**0.000**	**0.000**	**0.017**	**0.044**	0.532	**0.009**
	AUC	F	1.503	0.784	**4.504**	**12.105**	**19.471**	**72.703**	**4.951**	5.025	1.187	**11.927**
		p	0.229	0.382	**0.041**	**0.001**	**0.000**	**0.000**	**0.030**	0.032	0.284	**0.002**

Table H3: ANOVA results for comparing parameter categories using the 54 models that incorporated SDM2's results according to the various evaluation criteria. F refers to the F-ratio and p refers to the p-value. Significant differences in sample means are highlighted in red. F_{crit} = 4.03.

			Parameters									
			Elevation Access (m)			Number of Dispersal Choices			Choosing Location	Maximum Dispersal Distance (km)		
			100 vs. 200	200 vs. 300	100 vs. 300	2 vs. 5	5 vs. 10	2 vs. 10	HS vs. HS DIS	100 vs. 200	200 vs. 300	100 vs. 300
Evaluation Criteria	Slope	F	**5.292**	0.459	**9.898**	3.417	**5.800**	**23.171**	**11.229**	**6.174**	0.551	**12.404**
		p	**0.028**	0.503	**0.003**	0.073	**0.022**	**0.000**	**0.002**	**0.018**	0.463	**0.001**
	R^2	F	0.107	0.105	0.482	0.622	**15.690**	**17.832**	**4.718**	**6.276**	0.146	**4.863**
		p	0.746	0.748	0.492	0.436	**0.000**	**0.000**	**0.034**	**0.017**	0.705	**0.034**
	Overall Accuracy	F	2.616	0.150	**4.265**	**7.176**	**9.116**	**30.422**	2.749	**14.413**	1.157	**26.236**
		p	0.115	0.701	**0.047**	**0.011**	**0.005**	**0.000**	0.103	**0.001**	0.290	**0.000**
	Kappa	F	1.633	0.111	2.748	**5.221**	3.018	**14.947**	0.454	**5.221**	3.018	**14.947**
		p	0.210	0.741	0.107	**0.029**	0.091	**0.000**	0.503	**0.029**	0.091	**0.000**
	TSS	F	0.569	0.022	0.795	**4.201**	0.055	3.326	0.541	**5.846**	0.871	**11.496**
		p	0.456	0.883	0.379	**0.048**	0.815	0.077	0.465	**0.021**	0.357	**0.002**
	AUC	F	0.905	0.164	1.665	**8.474**	**10.280**	**33.572**	8.009	0.820	0.077	1.503
		p	0.348	0.688	0.206	**0.006**	**0.003**	**0.000**	0.007	0.372	0.782	0.229

b.) Box plots

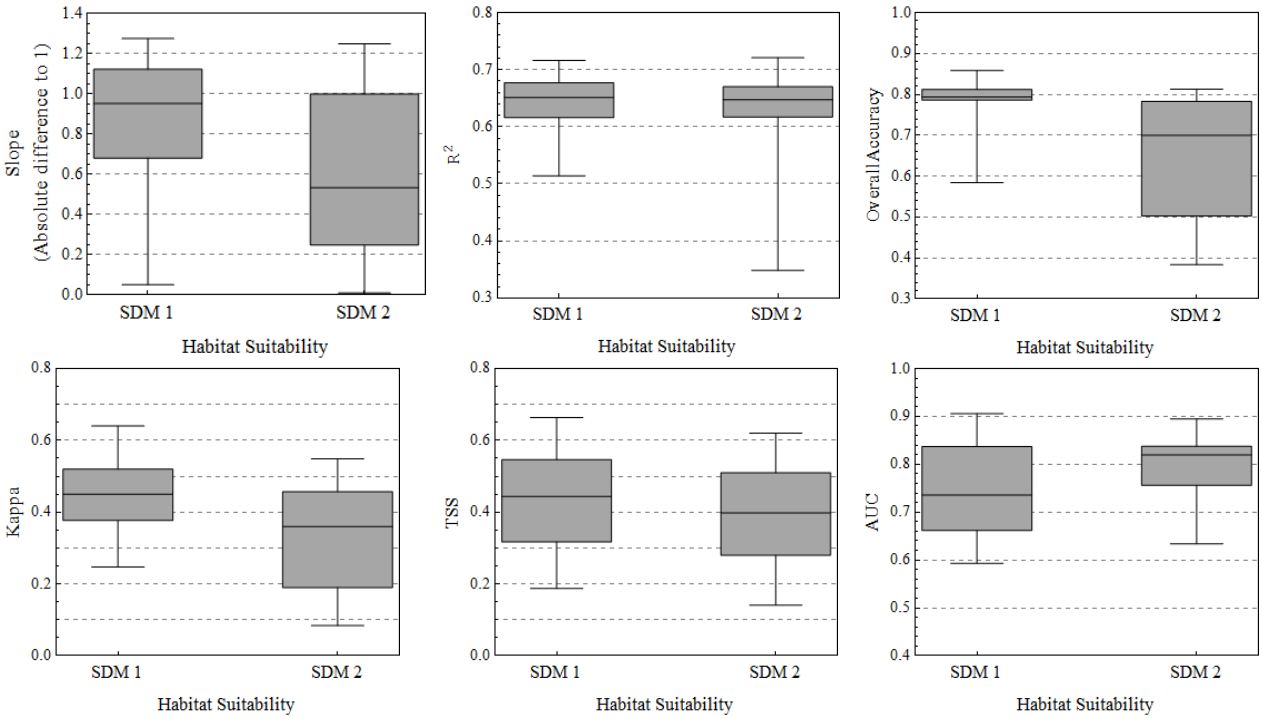


Figure H1: Box plot results for varying the habitat suitability parameter according to the various evaluation criteria.

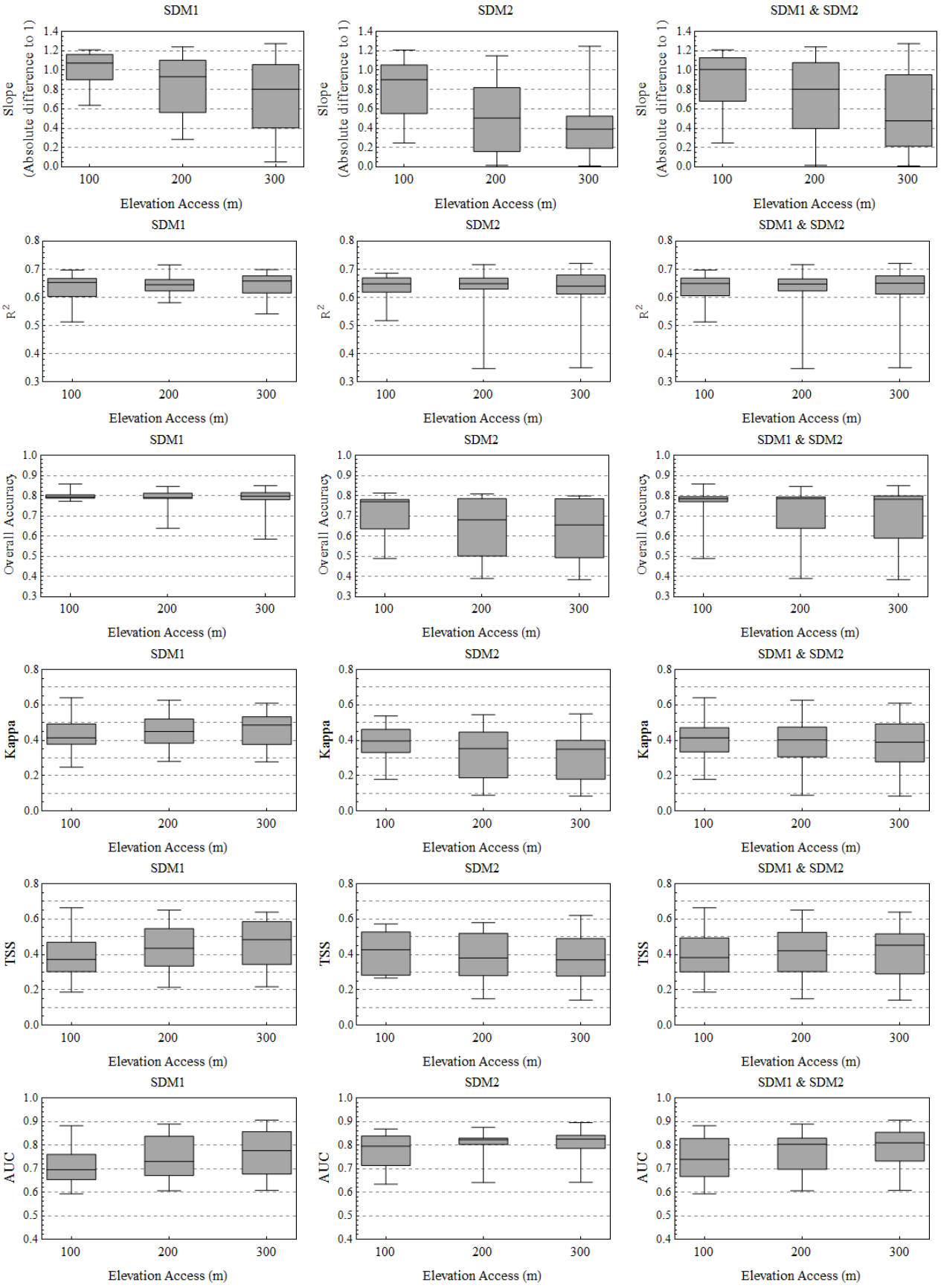


Figure H2: Box plot results for varying the elevation access parameter according to the various evaluation criteria.

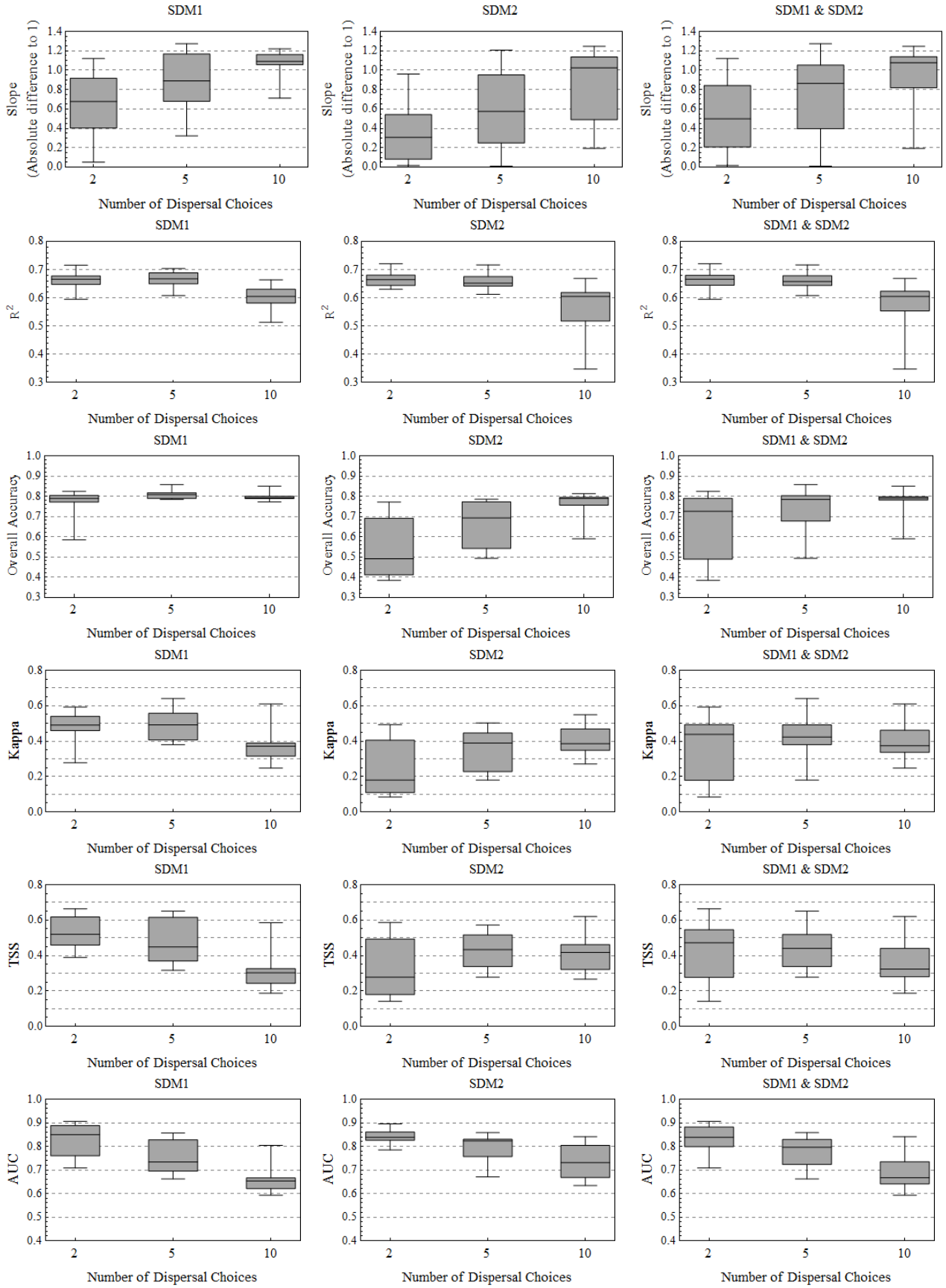
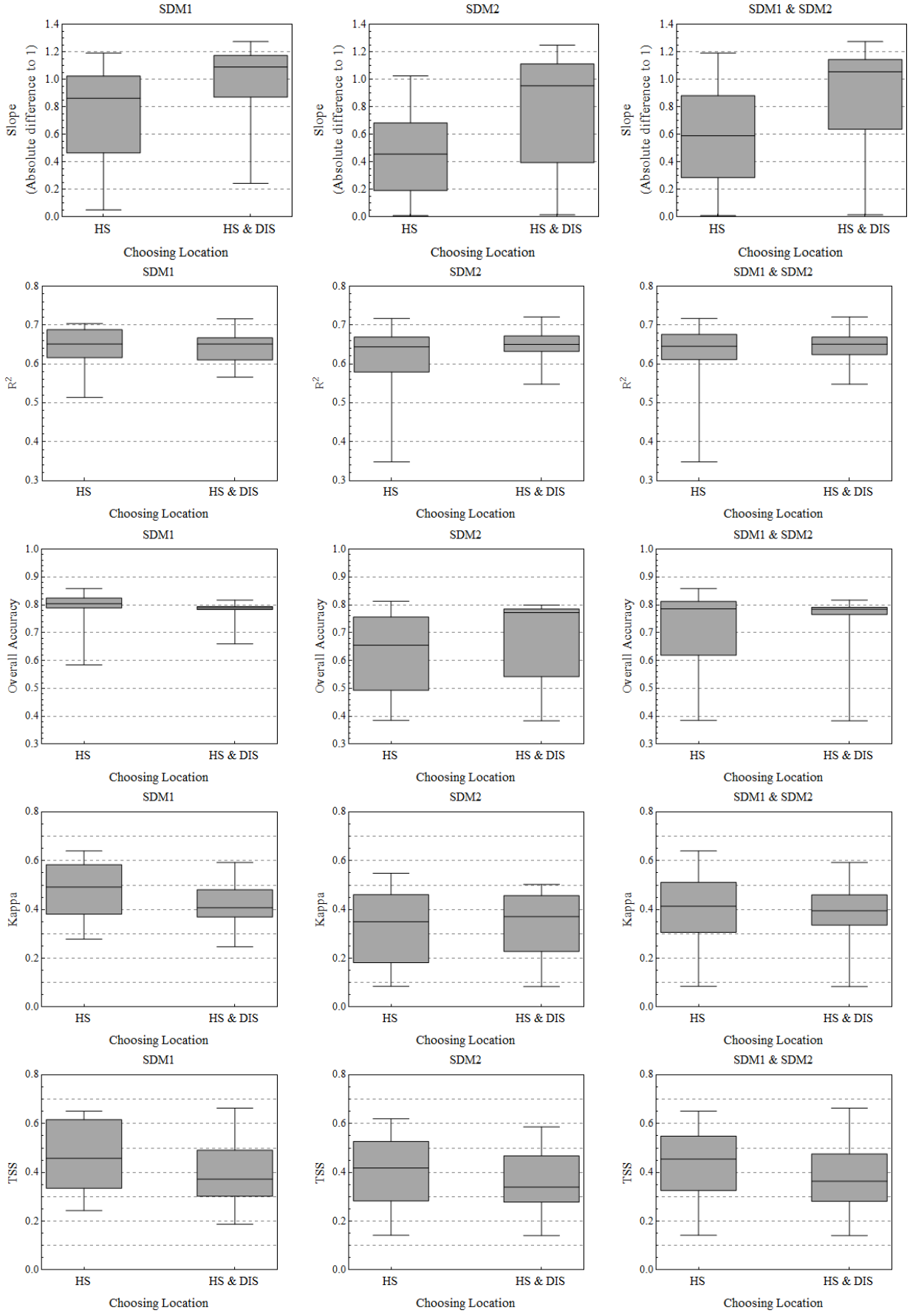


Figure H3: Box plot results for varying the dispersal choices parameter according to the various evaluation criteria.



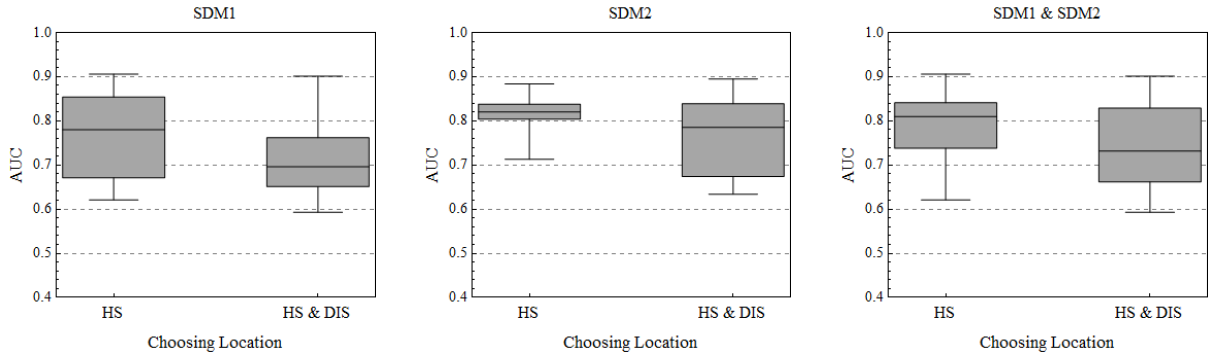
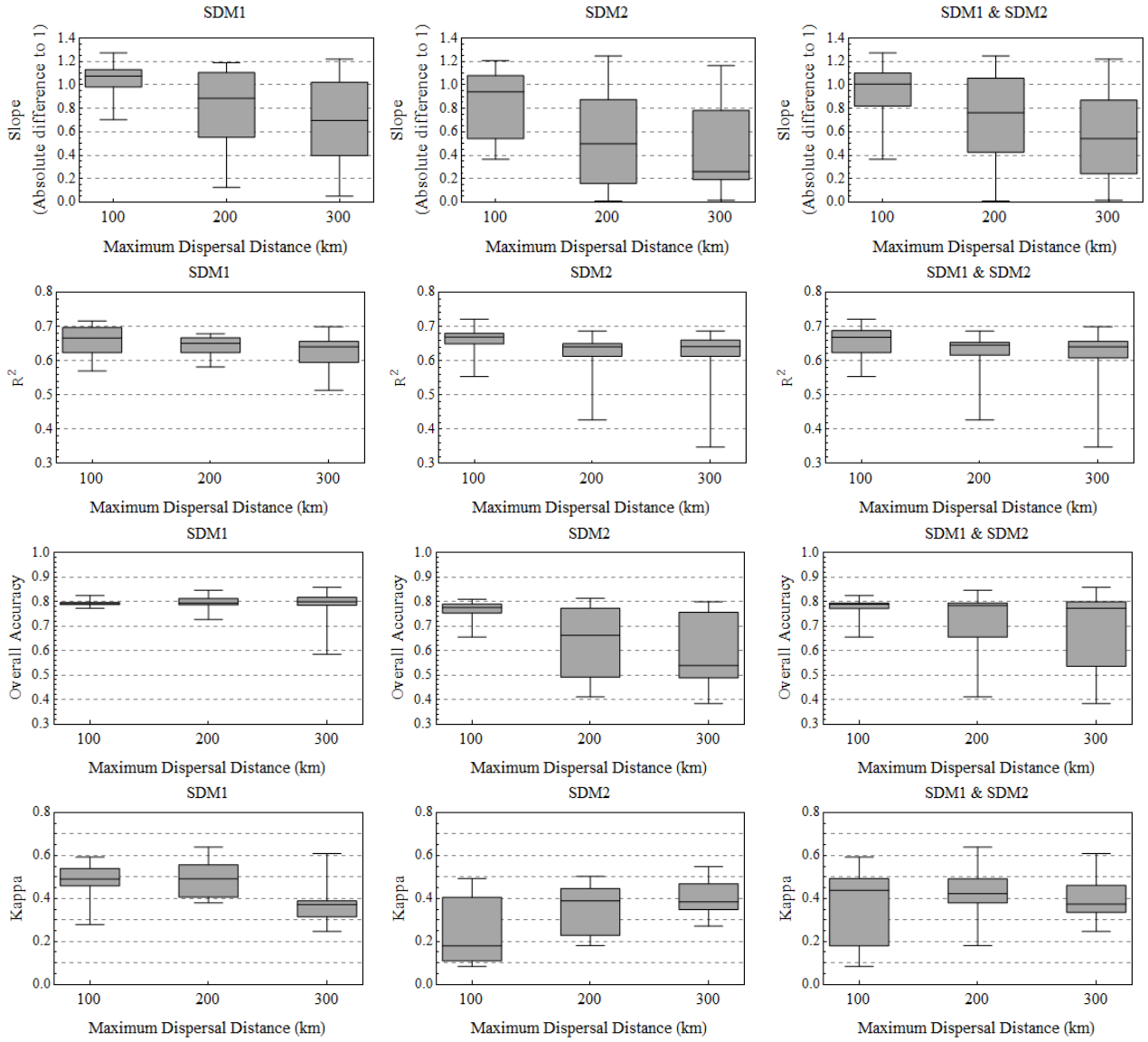


Figure H4: Box plot results for varying the dispersal function parameter according to the various evaluation criteria.



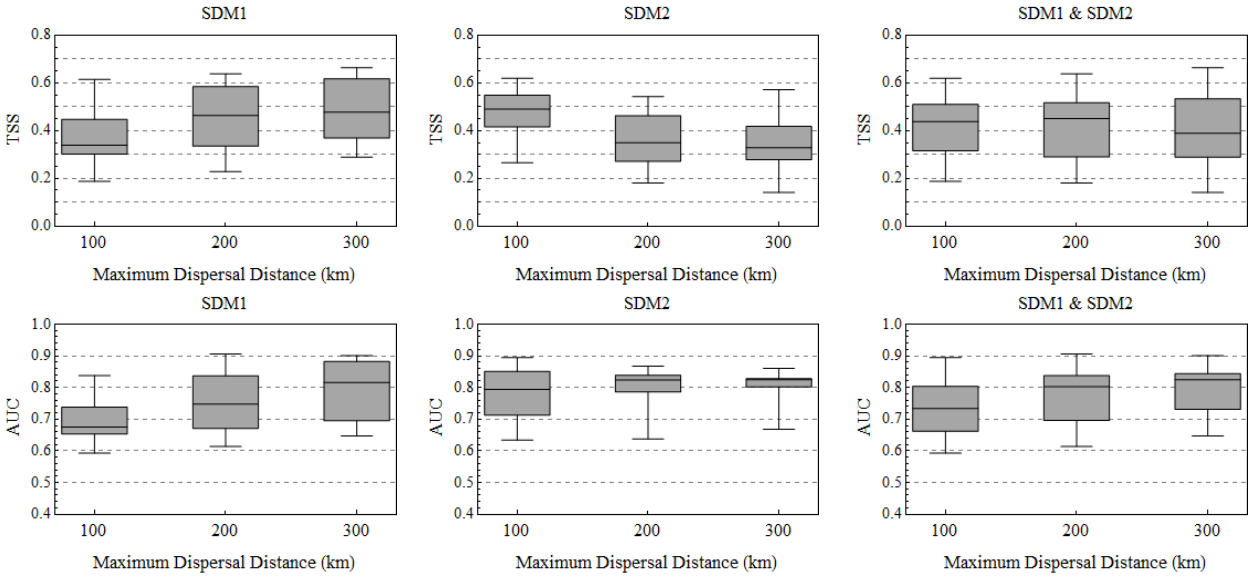


Figure H5: Box plot results for varying the dispersal distance parameter according to the various evaluation criteria.